Lecture notes for the course

V5B2 - Selected topics in Analysis and PDE -

Geometric Optimal Control

at the University of Bonn, Winter term 2018-19

Illia Karabash

These are shortened lecture notes, which may contain misprints. Corrections are welcome and should be sent to:

ikarabas(at)uni-bonn(dot)de or to i(dot)m(dot)karabash(at)gmail(dot)com

These notes cannot substitute textbooks and monographs listed below.

Books and lecture notes

[AS] A.A. Agrachev, Y. Sachkov, Control theory from the geometric viewpoint. Springer, 2013.

[BC] M. Bardi, I. Capuzzo-Dolcetta, Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations. Springer, 2008.

[B] J.F. Bonnans, Course on Optimal Control "Part I: the Pontryagin approach", http://www.cmap.polytechnique.fr/ bonnans/notes/oc/ocbook.pdf

[CS] P. Cannarsa, C. Sinestrari, Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004.

[E] L.C. Evans, Lecture notes of the course "An Introduction to Mathematical Optimal Control Theory", https://math.berkeley.edu/ evans/control.course.pdf

[FR] W.H. Fleming, R.W. Rishel, Deterministic and stochastic optimal control (Vol. 1). Springer Science & Business Media, 2012.

[GS] I.I. Gihman, A.V. Skorohod, Controlled stochastic processes. Springer Science & Business Medi, 2012.

[LM] E.B. Lee, L. Markus, Foundations of optimal control theory. Wiley, 1967.

[SL] H. Schättler, U. Ledzewicz, Geometric optimal control: theory, methods and examples. Springer, 2012.

[O] B. Øksendal, Stochastic differential equations. Springer, 2003.

Contents

1 Overview. Main notions and an example.

1.1 Example (minimum-time control of a harmonic oscillator).

Let $y(t) \in \mathbb{R}$ be the vertical coordinate of a unit mass hanging from a spring and subjected to external force $u(t) \in \mathbb{R}$ depending on time $t \in \mathbb{R}$. Then

$$y''(t) = -y(t) + u(t) \qquad \text{(here and below } y' = \partial_t y = \frac{\mathrm{d}y}{\mathrm{d}t}\text{)}, \tag{*}$$

$$\text{the initial position } y(t_0) = y^{[0]} \text{ and speed } y'(t_0) = y^{[1]} \text{ are given.} \tag{**}$$

The force satisfies the constraints $|u(t)| \leq 1$. When $u(t) \equiv 0$, the only equilibrium position is $y = 0$.

Problem. Our goal is to design the external force function $u(\cdot)$ that brings the motion in to a stop at $y = 0$ in the minimum possible time.

1.2 Terminology and the rigorous statement of the minimum time problem.

Let us rewrite the problem via a system of 1st order differential equations (eq-s) and pose it rigorously.

Let

$$x(t) = x = (x_1; x_2)^\top = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2,$$

where $x_1(t) = y(t)$, $x_2(t) = y'(t)$. Then $(*)$-$(**)$ becomes

$$x'(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t) \tag{CS}$$

$$x(t_0) = x^{[0]}, \quad x^0 = \begin{pmatrix} x_1^0(t_0) \\ x_2^0(t_0) \end{pmatrix} = \begin{pmatrix} y^{[0]} \\ y^{[1]} \end{pmatrix}.$$

In the sequel, we will write x^0 instead of $x^{[0]}$ for the sake of brevity.

Terminology.

(i) The vector-valued differential eq. (CS) is called *a control system*.

(ii) $x(t)$ is *the state* of the control system at time t,

(iii) x^0 is called *an initial state*.

(iv) The function $u(t)$, $t \in (t_0, +\infty)$ is called *a control strategy*. It is always assumed to be measurable. A simplified (but non-rigorous for some purposes) version of the measurability assumption is that

$$u(\cdot) \text{ is a piecewise continuous function on any finite interval } [t_0, t].$$

A control strategy is also assumed to satisfy the control constraint $u(t) \in U$, where $U = [-1, 1]$ is *the control set*.

Remark. The assumptions of (iv) means that the family of control strategies is

$$L_U^\infty(t_0,+\infty) := \{u \in L^\infty(t_0,+\infty) : u(t) \in U \text{ for almost all (a.a.)} t \in (t_0,+\infty)\}.$$

This family is called *the class of admissible controls*. Sometimes U is unbounded, and then, wider classes of admissible controls are considered, e.g.,

$$L^1_{\mathrm{loc},U}(t_0,+\infty) := \{\text{measurable } u(\cdot) \text{ on } (t_0,+\infty) : u \in L^1_U(t_0,t) \text{ for all } t \in (t_0,+\infty)\}.$$

The solution $x(\cdot)$ to the initial value problem for (CS) is called *the trajectory* of the system and is denoted by $x^{t_0,x_0,u}(\cdot)$.

Let

$$\tau(x_0,u) := \min\{t \geq t_0 : x^{t_0,x_0,u}(t) = 0\}$$

be *the exit time* of $x^{t_0,x_0,u}(\cdot)$, i.e., the first time when the trajectory reaches *the target set*, which in our example consist of one point $\{0\}$.

Convention. If a set S is empty, $S = \varnothing$, then $\min S = \inf S := +\infty$.

So $\tau(x_0,u) = +\infty$ if $x^{t_0,x^0,u}(t) \neq 0$ for all $t \geq t_0$.

Terminology. We denote by $\mathcal{C}_{[t_0,+\infty)}(0)$ the set of $x^0 \in \mathbb{R}^2$ such that (s.t.) $\tau(x_0,u) < +\infty$ for a certain control strategy $u(\cdot)$. That is, $\mathcal{C}_{[t_0,+\infty)}(0)$ is *the set of points controllable to* $\{0\}$ by a certain control strategy $u(\cdot)$ in the sense that $x(t) = 0$ at a certain finite moment $t \in [t_0,+\infty)$.

Minimum time problem for the harmonic oscillator. Given $x^0 \in \mathcal{C}_{[t_0,+\infty)}(0)$,

$$\text{minimize } \tau(x^0,u) \text{ over all } u(\cdot) \in L_U^\infty(t_0,+\infty). \tag{MTP}$$

(MTP) means that

- we have to find the value $T(x^0)$ at $x = x^0$ of the minimum time function

$$T(x) := \inf\{\tau(x,u) : u \in L_U^\infty(t_0,+\infty)\}$$

- and find all $u(\cdot)$ achieving the minimum in (MTP), i.e., all $u \in L_U^\infty(t_0,+\infty)$ s.t. $\tau(x^0,u) = T(x^0)$ (or, equivalently, s.t. $x^{t_0,x^0,u}(T(x^0)) = 0$).

Terminology. The corresponding $u(\cdot)$ and the trajectory $x^{t_0,x^0,u}(\cdot)$ are called *time-optimal*.

Quite general existence theorems are available.

In particular, for the harmonic oscillator (MTP), $\mathcal{C}_{[t_0,+\infty)}(0) = \mathbb{R}^2$ and there exist a time-optimal control for every $x^0 \in \mathbb{R}^2$. This can be shown with the use using Maximum Principle or Pontryagin Maximum Principle (PMP).

1.3 General problem in the Bolza form.

Consider a control system

$$x'(t) = f(x(t), u(t)), \quad t \in [t_0, T], \tag{CS}$$

where the dynamics evolves in the state space $\mathbb{R}^n$, i.e., $x(t) = (x_1(t); \dots; x_n(t))^\top \in \mathbb{R}^n$.

For each t, we assume $u(t) \in U$ with the control set $U \subset \mathbb{R}^m$. The class of admissible controls is

$$L^\infty_{\mathrm{loc},U}(t_0, +\infty) := \{u(\cdot) : u \in L^\infty_U(t_0, t) \quad \forall \text{ finite } t > t_0\}.$$

If you are not familiar with L^p-spaces and the Lebesgue measure, you can replace in all the statements of this section $L^\infty_{\mathrm{loc},U}(t_0, +\infty)$ by the class of admissible controls consisting of

$$\text{piecewise continuous functions bounded on each bounded interval.} \tag{SA}$$

Remark. If $u(\cdot) \equiv \widetilde{u}$ is constant, then (CS) is a dynamical system. Assume U is a finite set $U = \{\widetilde{u}^1, \widetilde{u}^2, \dots, \widetilde{u}^{\mathcal{N}}\}$. Then the control system is a collection of dynamical systems $x'(t) = f(x, \widetilde{u}^j)$ defined by vector-fields $f(x, \widetilde{u}^j)$ (in the sense that we can switch between them to achieve a desired result). The case when U is an infinite set is a generalization of this situation. In most of practical situations, U is compact and convex.

Presently, we impose no assumptions on U.

Assume that:

(A1) $f \in C_{\mathrm{loc}}(\mathbb{R}^n \times \mathbb{R}^m)$ (this means that f is continuous in all balls in $\mathbb{R}^n \times \mathbb{R}^m$, but we do not assume it uniformly bounded in $\mathbb{R}^n \times \mathbb{R}^m$),

(A2) $\partial_x f(x, u) := \frac{\partial f}{\partial x}$ exists for all pairs $(x; u) \in \mathbb{R}^n \times \mathbb{R}^m$,

(A3) $\partial_x f(x, u) \in C_{\mathrm{loc}}(\mathbb{R}^n \times \mathbb{R}^m)$.

Theorem (see e.g. [CS, SL])**.** *Let u be admissible, i.e., $u \in L^\infty_{\mathrm{loc},U}(-\infty, +\infty)$. Let f satisfies (A1)-(A3). Let $x^0 \in \mathbb{R}^n$ be the initial value of the state,*

$$x(t_0) = x^0 \tag{IS}$$

Then there exists $\delta > 0$ such that the initial value problem (CS), (IS) has a unique solution $x(t) = x^{t_0, x^0, u}(t)$ in the interval $t \in [t_0, t_0 + \delta)$ in the Carathéodory sense, i.e., in the sense that

$$x(t) = x^0 + \int_{t_0}^{t} f(x(s), u(s)) \mathrm{d}s, \quad t \in [t_0, t_0 + \delta).$$

This solution can be extended to a maximal interval of existence $(\tau_-, \tau_+) \subset \mathbb{R}$.

Remark.

(i) It is possible that $\tau_+ < +\infty$. This can happen when $\lim_{t \to \tau_+} x(t) = \infty$.

(ii) $x(\cdot)$ is locally absolutely continuous in (τ_+, τ_-), $x \in AC_{\mathrm{loc}}(\tau_+, \tau_-)$, and the equality (CS) holds a.e. on (τ_-, τ_+).

(iii) For a simplified version of the theorem and the above remarks (i)-(ii), one can take u piecewise-continuous and locally bounded. Then equality (CS) is satisfied everywhere on (τ_-, τ_+) except a finite number of points, the solution is continuous and piecewise-differentiable.

Definition (Admissible trajectory)**.** Let $u(\cdot)$ be admissible, i.e., $u \in L^\infty_{\mathrm{loc},U}(t_0, +\infty)$. Let $x(\cdot)$ be the unique solution to (CS), (IS) defined on a maximal interval of its existence. Then we say that a pair $(u; x)$ is an *admissible control-trajectory pair*.

The optimal control problem consists of minimization of a certain cost function over the family of all admissible control-trajectory pairs.

Definition. Consider the cost function (or objective) in *the Bolza form*

$$J(u) = \int_{t_0}^{T} \ell(x(s), u(s))\mathrm{d}s + \varphi(x(T)).$$

Here $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, which maps $(x; u) \in \mathbb{R}^n \times \mathbb{R}^m$ to $\ell(x, u) \in \mathbb{R}$, is called *running cost*. The function $\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is called *the terminal cost* (or penalty term); T is called *the terminal time*.

We assume that

(A4) ℓ satisfies (A1)-(A3),

(A5) $\varphi \in C^1_{\mathrm{loc}}(\mathbb{R} \times \mathbb{R}^n)$.

The terminal time T and the corresponding terminal state $x(T)$ are assumed to satisfy *terminal constraints*

$$(T; x(T)) \in N, \text{ where } N = \{(t; x) \in \mathbb{R} \times \mathbb{R}^n : \Psi(t, x) = 0\}$$

with a certain Ψ, which maps $(t; x) \in \mathbb{R} \times \mathbb{R}^n$ to $(\Psi_0(t, x); \ldots; \Psi_{n-k}(t, x))^\top \in \mathbb{R}^{n+1-k}$, $0 \le k \le n$.

The terminal set N and so the function Ψ usually are of quite simple forms.

However, to be flexible, let us assume that:

(A6) N is a k-dimensional embedded C^1-manifold in $\mathbb{R} \times \mathbb{R}^n$.

This means that $\Psi \in C^1_{\mathrm{loc}}(\mathbb{R} \times \mathbb{R}^n)$ and its Jacobian matrix

$$D\Psi := \begin{pmatrix} \partial_t \Psi_0 & \partial_{x_1} \Psi_0 & \ldots & \partial_{x_n} \Psi_0 \\ \ldots & \ldots & \ldots & \ldots \\ \partial_t \Psi_{n-k} & \partial_{x_1} \Psi_{n-k} & \ldots & \partial_{x_n} \Psi_{n-k} \end{pmatrix}$$

is of full rank on N, i.e.,

$$\operatorname{rank} D\Psi(t, x) = n - k + 1 \text{ for all } (t; x) \text{ such that } \Psi(t; x) = 0.$$

Optimal control problem (OCP). Given the initial state $x^0 \in \mathbb{R}^n$, minimize the cost function $J(u)$ over all admissible control-trajectory pairs $(u; x)$ and intervals $[t_0, T]$ connected by the terminal constraint $(T, x(T)) \in N$.

Examples of various types of problems and terminal constraints.

(i) Let $k = 1$ and $\Psi(t, x) = x - x^{target}$ with a certain target point $x^{target} \in \mathbb{R}^n$. Let $\ell(x; u) \equiv 1$ and $\varphi \equiv 0$. Then we obtain the problem of *minimum time control to the target point* $\{x^{target}\}$.

(ii) If $\Psi(t, x) = \Psi(x)$, i.e., Ψ is independent of time t, (OCP) is said to be with free terminal time. This is the case for (i).

(iii) Let $T > t_0$ be fixed and $\Psi_0(t, x) = t - T$. This is the problem with *fixed terminal time*. Note that it is possible that other coordinates of Ψ (that is, $\Psi_1(t, x)$, $\dots$,) define additional constraints on the terminal state $x(T)$.

Remark.

(i) (CS) and the cost functional $J(t)$ are time-invariant because the functions f, ℓ, and φ do not depend on t explicitly. If additionally the problem has free terminal time, i.e., $\Psi(t, x) = \Psi(x)$, then (OCP) is time-invariant and its solutions do not change essentially under the shifts of time.

(ii) More general control systems have the form

$$x' = f(t, x, u)$$

or may have ℓ or φ time-dependent.

References for Section 1.

[BC] M. Bardi, I. Capuzzo-Dolcetta, Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations. Springer, 2008.

[CS] Cannarsa, P., Sinestrari, C., Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004.

[SL] H. Schättler, U. Ledzewicz, Geometric optimal control: theory, methods and examples. Springer, 2012.

2 Overview (continuation). Hamilton-Jacobi-Bellman (HJB) equations and Maximum Principle.

2.1 HJB equations and the value function.

The HJB equation is something like a sufficient condition for optimality of control strategies (in the sequel, for brevity, we will sometimes call control strategies simply controls).

If we change the initial state x^0 or the initial time t_0 the value of J changes even if the control strategy u is the same. So, to consider problems (CS), (IS) for various x^0, we have to write the cost functional as $J(u; t_0, x^0) = J_{t_0,x^0}(u)$.

Definition. The function $V : \mathbb{R}^n \to (-\infty, +\infty]$ defined by

$$V(t_0, x^0) = \inf_{u \in L^\infty_{\mathrm{loc},U}(\mathbb{R})} J_{t_0,x^0}(u)$$

is called the *value function* of (OCP).

Under certain (very restrictive) assumptions, V satisfies the HJB differential equation:

$$\partial_t V(t,x) + \min_{\widetilde{u} \in U} (\partial_x V(t,x) \cdot f(x, \widetilde{u}) + \ell(x, \widetilde{u})) = 0,$$

where $\partial_x V = (\partial_{x_1} V, \dots, \partial_{x_n} V)$ can be perceived as a row-vector and the scalar product of two vectors $\partial_x V(t,x) \cdot f(x, \widetilde{u})$ can be perceived as a product of a row-vector and a column-vector.

In very general setting, V is a unique solution to this equation in a special generalized sense (e.g., a unique viscosity or proximal solution, see [CS, CV03]).

If we have solved the HJB equation, the value of $V(t_0, x^0)$ gives the sufficient condition of optimality in the following sense: if for a certain $u(\cdot)$, we have $J_{t_0,x^0}(u) = V(t_0, x^0)$, then $u(\cdot)$ is optimal.

For the minimum time control problem to a target set, the HJB eq. takes the form (see [BC, CS])

$$\min_{\widetilde{u} \in U} (\partial_x V(x) \cdot f(x, \widetilde{u})) = -1 \qquad (\text{ or } \min_{\widetilde{u} \in U} \nabla_{f(x,\widetilde{u})} V(x) = -1 \ ),$$

where $\nabla_w V(x) := \langle \partial_x V(x), w \rangle_{\mathbb{R}^n}$ is the directional derivative in the direction $w \in \mathbb{R}^n$.

2.2 Maximum Principle for time-invariant f and ℓ.

It seems that this was Vladimir Boltyansky who gave the name of Lev Pontryagin to Maximum Principle around 1960. Pontryagin Maximum Principle (PMP) is a first-order necessary condition of optimality. About the history of PMP, it is possible to read in [PP12].

We assume that (CS) and J are as above, i.e., time-invariant.

Definition. The control Hamiltonian function $H : [0, +\infty) \times (\mathbb{R}^n)^\top \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ for (OCP) is is defined by

$$H(\lambda_0, \lambda, x, u) = \lambda_0 \ell(x,u) + \lambda f(x,u).$$

Here $\lambda = (\lambda_1, \dots, \lambda_n) \in (\mathbb{R}^n)^\top$ is a row-vector (or covector) following the notation of [SL].

Terminology. The pair $(\lambda_0; \lambda)$ is called a *multiplier*, λ is called sometimes a *cotangent vector*. The covector-valued function $\lambda(t)$ satisfying the conditions of PMP given below is called *an adjoint variable*.

Theorem (PMP, see [SL])**.** *Let the control-trajectory pair $(u_*(t); x_*(t))$, $t \in [t_0, T]$, be optimal. Then there exists $\lambda_0 \in [0, +\infty)$ and a covector-valued function $\lambda : [t_0, T] \to (\mathbb{R}^n)^\top$ such that:*

(i) $(\lambda_0; \lambda(t)) \neq 0$ for all t as an element of the vector space $\mathbb{R} \times (\mathbb{R}^n)^\top$ (nontriviality of the multiplier)

(ii) the adjoint variable λ is a solution to the linear (system of) ODE

$$\lambda'(t) = -\lambda_0 \partial_x \ell(x_*(t), u_*(t)) - \lambda(t) \partial_x f(x_*(t), u_*(t)) \tag{AdjEq}$$

(the adjoint equation).

(iii) There exists a constant C such that, for all t,

$$C = H(\lambda_0, \lambda(t), x_*(t), u_*(t)) = \min_{\widetilde{u} \in U} H(\lambda_0, \lambda(t), x_*(t), \widetilde{u})$$

(the minimum condition)

(iv) At the endpoint, the covector $(C\ ;\ -\lambda(T) + \lambda_0 \partial_x \varphi(x_*(T)))$ is orthogonal to the terminal manifold N, i.e., there exists $v \in (\mathbb{R}^{n+1-k})^*$ s. t.

$$C + v \partial_t \Psi(T, x_*(T)) = 0, \quad \lambda = \lambda_0 \partial_x \phi(x_*(T)) + v \partial_x \Psi(T, x_*(T)).$$

Definition. A contol-trajectory pair $(u(\cdot); x(\cdot))$ satisfying (i)-(iv) of PMP with a certain multiplier $(\lambda_0; \lambda(\cdot))$ is called an *extremal*. The corresponding 4-tuple $(u(\cdot); x(\cdot); \lambda_0; \lambda(\cdot))$ is called *an extremal lift*. This extremal lift is called normal if $\lambda_0 > 0$ and abnormal if $\lambda_0 = 0$.

Remark. It is possible that an extremal $(u; x)$ is a part of a normal extremal lift $(u; x; \lambda_0; \lambda)$ and an abnormal extremal lift $(u; x; 0; \widetilde{\lambda})$ simultaneously. An extremal (u, x) such that every associated extremal lift is $(u; x; \lambda_0; \lambda)$ is abnormal, i.e., has $\lambda_0 = 0$, are called *strictly abnormal*.

2.3 Bang-bang controls.

Definition. Assume that the control set is the interval $U = [u_-, u_+]$, $u_\pm \in \mathbb{R}$. A control strategy $u(\cdot)$ is called bang-bang if, on any finite interval $[t_0, T]$,

(i) $u(\cdot)$ is piecewise-constant with a finite number of points of discontinuity (after a possible correction on a set of measure 0),

(ii) on each intervals of constancy $u(t)$ takes one of the two extreme values $u_\pm$.

Time-points t where $u(t-0) \neq u(t+0)$ (the limit from the left is not equal to the limit from the right) are called switching time-point (we assume that the correction of (i) already have been done).

2.4 Chattering arcs and singular arcs.

Sometimes the structure of controls is not so good as it is described above for the bang-bang controls. Among the effects that can make control "not so good" are:

- *chattering arcs*, when control switches infinite number of times on a bounded time-interval
- *singular arcs*, when trajectory of $(\lambda; x)$ in $\mathbb{R}^n \times \mathbb{R}^n$ goes along the zero surface of switching function.

A famous example (see [SL]) of an optimal control with chattering arcs is given by the Fuller problem of the control of

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} x_2 \\ u \end{pmatrix} \quad \text{to the target point} \quad x^{target} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

with the cost function

$$J(u) = \int_{t_0}^{T} (x_1(t))^2 \mathrm{d}t.$$

To clarify the notion of singular arc, let us consider the following important type of control systems.

2.5 Minimum-time control of control-affine system.

Consider the control system

$$x'(t) = F(x(t)) + u(t)G(x(t)), \quad F, G : \mathbb{R}^n \to \mathbb{R}^n, \quad u(t) \in U = [-1, +1]. \tag{CAS}$$

The systems of such type are called time-invariant, single-input, control-affine systems.

Consider a minimum-time problem with a certain $x^{target} \in \mathbb{R}^2$.

PMP implies that every optimal pair $(u_*; x_*)$ satisfies

$$u_*(t) = -\operatorname{sgn}\langle \lambda(t), G(x_*(t))\rangle_{\mathbb{R}^n} \text{ for all } t \text{ such that } \langle \lambda(t), G(x_*(t))\rangle_{\mathbb{R}^n} \neq 0$$

Definition (see Section 2.8 in [SL]).

(i) The function $\Phi_{(u;x)}(t) := \langle \lambda(t), G(x(t))\rangle_{\mathbb{R}^n}$ is called *switching function* for the extremal pair $(u; x)$.

(ii) The extremal control strategy $u(\cdot)$ is called *singular* on an interval $[t_1, t_2]$ if $\Phi_{(u;x)}(t) \equiv 0$ on $[t_1, t_2]$.

References for Section 2.

[BC] M. Bardi, I. Capuzzo-Dolcetta, Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations. Springer, 2008.

[CS] Cannarsa, P., Sinestrari, C., Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004.

[CV03] I. Chryssochoos, R.B. Vinter, Optimal control problems on manifolds: a dynamic programming approach. Journal of mathematical analysis and applications 287(1), (2003), 118–140.

[PP12] Pesch, H.J., and Plail, M. "The cold war and the maximum principle of optimal control." Optimization Stories. Documenta Mathematica (2012).

[SL] H. Schättler, U. Ledzewicz, Geometric optimal control: theory, methods and examples. Springer, 2012.

3 Overview (continuation). Optimal synthesis, feedback, and control problems on manifolds.

3.1 Maximum Principle for minimum-time problems (MTP).

Corollary. *If the terminal constraint $N = \{(t,x) : \Psi(x) = 0\}$ do not depend on t, then $C = 0$ in the minimum condition (iii). That is, the minimum condition takes the form:*

$$0 \equiv H(\lambda_0, \lambda(t), x_*(t), u_*(t)) = \min_{\widetilde{u}\in U} H(\lambda_0, \lambda(t), x_*(t), \widetilde{u})$$

Proof. This follows directly from the terminal condition (iv) of PMP. □

Minimum time problem (MTP). Given the initial state $x(t_0) = x^0$ and the target state x^{target}, bring (CS) from x^0 to x^{target} in minimal possible time $T - t_0$ (where T is such that $x(T) = x^{target}$).

Theorem (PMP for MTP)**.** *Let the control-trajectory pair $(u_*(t); x_*(t))$, $t \in [t_0, T]$, be a minimizer of MTP. Then there exists $\lambda_0 \in [0, +\infty)$ and a covector-valued function $\lambda : [t_0, T] \to (\mathbb{R}^n)^\top$ such that:*

(i) $(\lambda_0; \lambda(t)) \neq 0$ for all t as an element of $\mathbb{R} \times (\mathbb{R}^n)^\top$ (nontriviality of the multiplier)

(ii) the adjoint variable λ is a solution to the linear (system of) ODE

$$\lambda'(t) = -\lambda_0 \partial_x \ell(x_*(t), u_*(t)) - \lambda(t)\partial_x f(x_*(t), u_*(t)) \qquad \text{(AdjEq)}$$

(the adjoint equation).

(iii) For all $t \in [t_0, T]$,

$$0 = H(\lambda_0, \lambda(t), x_*(t), u_*(t)) = \min_{\widetilde{u}\in U} H(\lambda_0, \lambda(t), x_*(t), \widetilde{u})$$

(the minimum condition).

Proof. For this problem the terminal condition (iv) of PMP gives only $0 \equiv H$ in (iii). □

Theorem (existence of optimizers of MTP)**.** *Assume that*

$$\text{the control set } U \text{ is compact;} \qquad \text{(H0)}$$

$$\text{there exists } K_1 > 0 \text{ such that } |f(x;u) - f(\widetilde{x}, u)| \leq K_1 |x - \widetilde{x}| \quad \forall\, x, \widetilde{x} \in \mathbb{R}^n,\, u \in U. \qquad \text{(H1)}$$

Assume also that the set $f(x, U) := \{f(x, u) : u \in U\}$ is convex for each $x \in \mathbb{R}^n$. Let $x^0 \in \mathcal{C}_{[t_0,+\infty)}$ (i.e., the state x^0 is controllable to x^{target} in a certain finite time). Then there exists an optimizer for MTP (i.e., a control-trajectory pair $(u_*(\cdot); x(\cdot))$ that minimize $J(u) = T - t_0$).

The proof will be given later following [CS].

3.2 Optimal synthesis for the controlled harmonic oscillator.

3.2.1 Maximum principle for the controlled harmonic oscillator

For the controlled harmonic oscillator, PMP leads to the following steps (see also [SL, Section 2.6.4]):

- We introduce a multiplier $(\lambda_0; \lambda(t))$, where $\lambda_0 = const \geq 0$ and the adjoint variable $\lambda = (\lambda_1; \lambda_2) : [t_0, T] \to (\mathbb{R}^2)^\top$ satisfies the **adjoint equation**

$$\lambda'(t) = -\lambda(t) \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

  This gives $\lambda_1' = \lambda_2$ and $\lambda_2'' = -\lambda_2$.

- We construct the control Hamiltonian

$$H(\lambda_0, \lambda, x, u) = \lambda_0 \ell + \lambda f = \lambda_0 + (\lambda_1 \ \lambda_2) \begin{pmatrix} x_2 \\ -x_1 + u \end{pmatrix}.$$

- **Maximum principle for minimum-time control problems**: Every optimal control-trajectory pair $(u_*(t); x(t))$ should satisfy

$$0 = H(\lambda_0, \lambda(t), x(t), u_*(t)) = \min_{\widetilde{u} \in U} H(\lambda_0, \lambda(t), x(t), \widetilde{u})$$

- The second equality (the minimal condition) gives $\lambda_2(t) u_*(t) = \min_{\widetilde{u} \in U} (\lambda_2(t) \widetilde{u})$. So

$$u_*(t) = -\operatorname{sgn} \lambda_2(t) \quad \text{if } \lambda_2(t) \neq 0.$$

  Here $\lambda_2(t)$ is a switching function.

- The first equality gives

$$0 = \lambda_0 + \lambda_2' x_2 + \lambda_2(-x_1 + u_*).$$

- The multiplier should be nondegenerate in the sense that $(\lambda_0; \lambda) \neq 0$ for all t. This implies that $\lambda_2(\cdot)$ is a nontrivial solution in the sense $\lambda_2(\cdot) \not\equiv 0$.

Lemma. *Let u be an optimal control for the harmonic oscillator. Then*

(i) u is bang-bang,

(ii) u switches between ± 1 exactly in time π.

Proof. Indeed, $u_*(t) = \operatorname{sgn} \lambda_2(t)$, $\lambda_2(\cdot) \not\equiv 0$ and its general form is $\lambda_2 = C_1 \sin(t + \theta_1)$ with $C_1 \neq 0$. □

Remark. Note that we have no additional information about C_1 and θ_1. However, $\lambda_2(t)$ is coupled with $x(t)$ by the equality $0 = \lambda_0 + \lambda_2' x_2 + \lambda_2(-x_1 + u_*)$ and the inequality $\lambda_0 \geq 0$. In the present case, we do not need this coupling to give a complete solution, but for other problems this coupling is useful.

For problems of optimal control to a target point x^{target}, it is easier to start from x^{target} running time backwards. It also makes sense to start from a special case of abnormal extremals.

3.2.2 Extremal synthesis for harmonic oscillator.

Consider trajectories of the extreme dynamical systems with $u = u_\pm = \pm 1$. They have the form

$$x_1(t) \mp 1 = -C_2 \cos(t - \theta_2), \quad x_2(t) = C_2 \sin(t - \theta_2).$$

Let us construct abnormal extremals going backward in time from the terminal time $T = 0$. An extremal is abnormal if $\lambda_0 = 0$. The equality $0 \equiv H$ gives

$$0 = \lambda_0 + \lambda_2' x_2 + \lambda_2(-x_1 + u_*(t)) = \lambda_2' x_2 + \lambda_2(-x_1 + u_*(t))$$

in the Carathéodory sense (in the present case, for every t except finite number of points in any finite interval). Taking it at $T = 0$, we get $\lambda_2(0) = 0$ since $|u(0 - 0)| = 1$. Here and below

$$u(t_1 \pm 0) \text{ are limits } \lim_{t \to 0 \pm 0} u(t_1 + t).$$

So the preceding zero of λ_2 is at $t = -\pi$, and switching for abnormal controls happens at times $t = -n\pi$, $n \in \mathbb{N}$.

One can see that the last part of an abnormal extremal trajectory without switches is either

$$\Gamma_+ : [-\pi, 0] \to \mathbb{R}^2 \text{ given by } x_1 = 1 - \cos t,\ x_2 = \sin t,$$

(corresponding to $u = 1$) or

$$\Gamma_- : [-\pi, 0] \to \mathbb{R}^2 \text{ given by } x_1 = -1 + \cos t,\ x_2 = -\sin t,$$

(corresponding to $u = -1$).

Then we switch control at $t = -\pi$ to another extreme value and continue the procedure switching u at times $t = -n\pi$.

There are exactly two abnormal extremal trajectories. Each consists of alternating semicircles with centers at $x = (\pm 1; 0)$. After each switch the radius growth by 2. These two extremal trajectories are strictly abnormal.

Now it is possible to construct the family $\mathbb{F}$ of all extremal trajectories x choosing arbitrary $t \in (-\pi, 0)$ as the last switching time. Such a description of all extremal trajectories is called *extremal synthesis*.

Switching between ± 1 occurs at the points of *switching locus* Υ,

$$\Upsilon := \{x : \operatorname{dist}(x, (2\mathbb{Z} + 1; 0)) = 1,\ x_1 x_2 \le 0\}, \text{ where } (2\mathbb{Z} + 1; 0) := \{(2n + 1; 0) : n \in \mathbb{Z}\}.$$

Let G_+ (G_-) be the set of points below (resp., above) Υ.

We have proved the following statement.

Lemma. *Let $(u_*(\cdot); x(\cdot))$ be an extremal control-trajectory pair. Then:*

(i) (after a possible correction of $u_*(\cdot)$ on a set of zero measure)

$$u_*(t) = \pm 1 \quad \text{if } x(t) \in \Gamma_\pm \cup G_\pm. \tag{FC}$$

(ii) Two different trajectories from the set $\mathbb{F}$ (of all extremal trajectories) do not intersect each other. Besides,

$$\mathbb{R}^2 \setminus \{0\} = \bigcup_{x \in \mathbb{F}} \{x(t) : t \in (-\infty, 0)\}$$

Terminology. Controls that are functions of states (as above) are called *feedback controls.*

Corollary. *Let $x^0 \neq 0$. Let $x \in \mathbb{F}$ be such that $x^0 \in \{x(t) : t \in (-\infty, 0)\}$. Let u_* be connected with x by (FC). Then (u_*, x) is the unique optimal control-trajectory pair for MTP with the initial state x^0.*

So, in this case, the extremal synthesis is also *optimal synthesis*, i.e., optimal (u_*, x) for all initial states $x^0 \in \mathcal{C}_{[t_0,+\infty)}$ were constructed from simpler pieces of dynamics with $u = const = u_\pm$.

3.3 Control systems with restricted state space and on smooth manifolds.

Let $M \subset \mathbb{R}^n$ be connected. Let us take M as a state space.

Definition (Admissible trajectory)**.** Let $u(\cdot)$ be admissible, i.e., $u \in L^\infty_{\text{loc},U}(-\infty, +\infty)$. Let $x(\cdot)$ be the unique solution to (CS) with the initial condition $x(t_0) = x^0 \in M$. Let (τ_-, τ_+) be a maximal open interval containing t_0 such that $x(\cdot)$ exists on (τ_-, τ_+) and $x(t) \in M$ for all $t \in (\tau_-, \tau_+)$. Then the pair $(u; x)$ on(τ_-, τ_+) is *admissible control-trajectory pair* in the state space M.

If the state space M is an open connected subset of $\mathbb{R}^n$, then PMP remains valid in the same form as it was given above.

Another important case is when $M \subset \mathbb{R}^n$ is a smooth manifold. One have to ensure that the dynamics of x stays on M. Let us consider a collection $V_u(x)$, $u \in \widetilde{U}$, of vector fields on M, that is, for each u,

$$V_u \text{ maps each } x \in M \text{ to } V_u(x) \in T_xM$$

(for simplicity we can assume that V_u is smooth for each value of control u).

Then the control system on the manifold can be defined as

$$x'(t) = V_u(x), \qquad x(t_0) = x^0 \in M.$$

3.3.1 Example 1. Dubins' car.

The car position is described by its coordinates $(x_1; x_2) \in \mathbb{R}^2$ (of the center of mass) and the angle θ of the car axis with positive x_1 axis. The car moves forward at velocity 1. We control only the steering $\theta' = u$ with the constraint $|u| \leq 1$. Then one gets the control system

$$x_1' = \cos\theta \tag{1}$$
$$x_2' = \sin\theta \tag{2}$$
$$\theta' = u \tag{3}$$

on the manifold $\mathbb{R}^2 \times S^1$, where S^1 is a unit circle. The solution is described partially in [BP03, Introduction] and completely in [SL96].

3.3.2 Example 2. Optimization of an optical resonator.

Optimization of resonators will be considered in more detail Section 11 following the papers [K13, KLV17] and the recent preprint [KKV18].

The problem of a minimization of the length of a 1-dimensional (1-dim.) optical cavity producing a given resonance $k \in \mathbb{C}_-$ can be reduced to minimum-time control of the system

$$x' = k(-x^2 + \varepsilon),$$

where $0 < \epsilon_1 \le \varepsilon(t) \le \epsilon_2$ and $x(t) \in \widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, i.e., x evolves on the Riemann sphere $\widehat{\mathbb{C}}$. The function $\varepsilon(\cdot)$ can be interpreted as a control. Physically it represents the layered structure of a nonhomogeneous optical cavity. Initial state $x^0 = -1$ and the target state $x^{target} = +1$ come from the radiation conditions.

References for Section 3.

[BP03] Boscain, U., Piccoli, B. (2003). Optimal syntheses for control systems on 2-D manifolds (Vol. 43). Springer Science Business Media.

[K13] I.M. Karabash, Optimization of quasi-normal eigenvalues for 1-D wave equations in inhomogeneous media; description of optimal structures, Asymptotic Analysis 81 (2013) no.3-4, 273-295.

[KLV17] Karabash, I. M., Logachova, O. M., Verbytskyi, I. V. Nonlinear Bang–Bang Eigenproblems and Optimization of Resonances in Layered Cavities. Integral Equations and Operator Theory 88(1),(2017), 15-44.

[KKV18] I.M. Karabash, H. Koch, I.V. Verbytskyi, Pareto optimization of resonances and minimum-time control, preprint arXiv:1808.09186, https://arxiv.org/pdf/1808.09186

[SL] H. Schättler, U. Ledzewicz, Geometric optimal control: theory, methods and examples. Springer, 2012.

[SL96] Soueres, P., Laumond, J. P. (1996). Shortest paths synthesis for a car-like robot. IEEE Transactions on Automatic Control, 41(5), 672-688.

4 Existence for Mayer's problem, compactness, and basic ODE results.

4.1 Optimal control problem in the Mayer form.

For the theory of the Mayer problem we mainly follow the monograph [CS].

Recall that we consider the control system

$$x'(t) = f(x(t), u(t)), \quad t \in [t_0, T]. \tag{CS}$$

We assume that $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous, and that the control strategy (or simply control) $u(\cdot)$ is measurable and for almost all (a.a.) t satisfies $u(t) \in U \subset \mathbb{R}^m$.

We will employ the following assumptions:

$$\text{the control set } U \text{ is compact;} \tag{H0}$$

$$\text{there exists } K_1 > 0 \text{ such that } |f(x;u) - f(\widetilde{x}, u)| \le K_1 |x - \widetilde{x}| \quad \forall\, x, \widetilde{x} \in \mathbb{R}^n,\, u \in U. \tag{H1}$$

So the family of admissible controls is $L^\infty_U(\mathbb{R})$.

Definition. Let $u(\cdot) \in L^\infty_U(\mathbb{R})$, let $I \subset \mathbb{R}$ be an interval. A function $x : I \to \mathbb{R}^n$ is called a solution to (CS) in the Carathéodory sense if, for any compact interval $[t_1, t_2] \subset I$, we have $x \in AC[t_1, t_2]$ and $x'(t) = f(x(t), u(t))$ for a.a. $t \in [t_1, t_2]$.

(H0) and (H1) imply the following uniform boundedness property for f:

$$|f(x, u)| \le C + K_1 |x| \text{ for all } x \in \mathbb{R}^n, \quad u \in U, \tag{UBf}$$

where $C = \max_{u \in U} |f(0, u)|$.

The arguments of standard ODE theory imply the global existence and uniqueness of solutions to (CS) equipped with the initial condition

$$x(t_0) = x^0, \tag{IS}$$

where $t_0 \in \mathbb{R}$ and $x^0 \in \mathbb{R}^n$.

Theorem (global existence and uniqueness)**.** *Assume (H0), (H1), and $u(\cdot) \in L^\infty_U(\mathbb{R})$. Then (CS), (IS) has a unique solution $x(t)$ (in the Carathéodory sense) defined for all $t \in \mathbb{R}$.*

We take this ODE result without proof.

Definition. By $x^{t_0, x^0, u}(\cdot)$ we denote the unique solution to (CS), (IS). Such solutions with admissible $u(\cdot)$ are called *(admissible) trajectories* of (CS).

Let $T > t_0$ be fixed. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be continuous.

Mayer's optimal control problem (MOCP).
Minimize $J(u) = \varphi(x^{t_0, x^0, u}(T))$ over all admissible control strategies $u \in L^\infty_U[t_0, T]$.

This is an example of a problem with *a fixed finite time horizon.*

It is a particular case of the Bolza optimal control problem (BOCP) with a fixed terminal time T. Recall that BOCP is to minimize the functional

$$J(u) = \int_{t_0}^{T} \ell(x(s), u(s))\mathrm{d}s + \varphi(x(T)).$$

If we take $\ell \equiv 0$, we get MOCP. Note that the terminal manifold $\{\Psi(t,x) = 0\}$ is defined by $\Psi = t - T$.

Remark. In the present case, T is a known fixed time. Therefore it is easy to reduce the (more general) Bolza problem to the (more special) Mayer problem. This will be done later.

Theorem. *Assume (H0), (H1), and $\varphi \in C_{\mathrm{loc}}(\mathbb{R}^n)$. Assume that*

$$\textit{the set } f(x,U) := \{f(x,u) : u \in U\} \textit{ is convex for each } x \in \mathbb{R}^n.$$

Then, for arbitrary $x^0 \in \mathbb{R}^n$, there exists an optimal control for MOCP.

The proof is in several steps.

The first step of the proof is to consider the minimizing sequence of control strategies $u_k(\cdot)$, i.e., $\{u_k\}_{k=1}^{\infty} \subset L_U^{\infty}[t_0,T]$ such that $\lim J(u_k) = \inf_{u \in L_U^{\infty}[t_0,T]} J(u)$. This means that for the corresponding trajectories $y_k(\cdot) := x^{t_0,x^0,u_k}(\cdot)$ we have

$$\lim_{k\to\infty} \varphi(y_k(T)) = \inf_{u \in L_U^{\infty}[t_0,T]} \varphi(x^{t_0,x^0,u}(T)).$$

We want to show that certain subsequence of $\{y_k\}$ converges uniformly to a function y that is also a trajectory, i.e., $y(\cdot) = x^{t_0,x^0,u_*}$ for certain $u_* \in L_U^{\infty}[t_0,T]$. Then, from continuity of $\varphi(\cdot)$, we obtain that y is an optimal trajectory and u_* is an optimal control.

The proof of the existence of a converging subsequence consists of several components:

- properties of trajectories under assumptions (H0), (H1),
- compactness property for the family of all trajectories $x(\cdot)$.

4.2 Properties of trajectories.

Let t_0, t_1 be fixed and such that $t_0 < t_1$.

Lemma (Uniform boundedness of trajectories x (UBx))**.** *Assume (H0), (H1). Then for any $r > 0$ there exists $R > 0$ such that*

$$|x^{t_0,x^0,u}(t)| \le R$$

for all $x^0 \in B_r := \{x \in \mathbb{R}^n : |x| < r\}$, $t \in [t_0,t_1]$, and $u \in L_U^{\infty}[t_0,t_1]$.

Gronwall's inequality. *Let $z \in AC[t_0,t_1]$, $z(t) \ge 0$ for all t, and*

$$z(t) \le k(t) + \int_{t_0}^{t} z(s)v(s)\mathrm{d}s, \quad t \in [t_0,t_1],$$

where $k \in C^1[t_0, t_1]$, $v \in C[t_0, t_1]$ are nonnegative functions. Then

$$z(t) \le k(t_0) e^{\int_{t_0}^t v(s)\mathrm{d}s} + \int_{t_0}^t k'(s) e^{\int_s^t v(r)\mathrm{d}r}, \qquad t \in [t_0, t_1]. \tag{GrIn}$$

In particular, if $k = const \ge 0$ and $v = const \ge 0$, we have

$$z(t) \le k e^{v(t-t_0)}, \quad t \in [t_0, t_1].$$

Proof of Lemma UBx. Let $x = x^{t_0, x^0, u}$. Then (H0), (H1), and (UBf) imply

$$|x(t)| \le |x^0| + \int_{t_0}^t (C + K_1 |x(s)|)\mathrm{d}s \le |x_0| + C(t_1 - t_0) + K_1 \int_{t_0}^t |x(s)|\mathrm{d}s.$$

(GrIn) applied to $|x|$ concludes the proof. □

Lemma (Uniform continuity w.r.t. initial state)**.** *Assume (H1). Then there exists $c = const > 0$ such that*

$$|x^{t_0, x^0, u}(t) - x^{t_0, x^1, u}(t)| \le c|x^0 - x^1|$$

for all $t \in [t_0, t_1]$, $x^0, x^1 \in \mathbb{R}^n$, and $u \in L^\infty_U[t_0, t_1]$.

Proof. It is enough to apply (GrIn) to the function $z = |x^{t_0, x^0, u} - x^{t_0, x^1, u}|$. □

4.3 Compactness property.

Theorem (compactness of trajectories)**.** *Assume (H0), (H1). Assume that*

$$\text{the set } f(x, U) := \{f(x, u) : u \in U\} \text{ is convex for each } x \in \mathbb{R}^n.$$

Let $y_k(t) := x^{t_0, x^k, u_k}(t)$, $t \in [t_0, t_1]$, be a sequence of trajectories with certain initial states $x^k \in \mathbb{R}^n$. Assume that $\{y_k(\cdot)\}$ is uniformly bounded on $[t_0, t_1]$, i.e.,

$$|y_k(t)| \le R \text{ for all } t \in [t_0, t_1] \text{ and } k.$$

Then there exists a subsequence $\{y_{k_\nu}\}$ and a trajectory $y = x^{t_0, x^k, u_k}$ such that $y_{k_\nu} \to y$ uniformly on $[t_0, t_1]$.

The proof will be given in the next lecture.

Remark. Without the assumption that $f(x, U)$ are convex, the compactness of trajectories may fail. Consider, the system $x' = u$, where $x \in \mathbb{R}$, $u \in U = \{-1, 1\}$. Consider controls $u_k(\cdot)$ alternating between ± 1 on intervals of length $1/k$ and put $x^k = 0$ for all k. Then $y_k \to 0$ uniformly. However, $y \equiv 0$ is not a trajectory since $y' \notin U$ for all t.

4.4 Proof of the existence of optimizer for Mayer's problem.

We take the minimizing sequences $\{u_k\}$ and $\{y_k\} = \{x^{t_0,x^0,u_k}\}$. They have the property

$$\lim_{k\to\infty} \varphi(y_k(T)) = \inf_{u\in L^\infty_U[t_0,T]} \varphi(x^{t_0,x^0,u}(T)).$$

By UBx, $\{y_k\}$ is uniformly bounded on $[t_0, T]$. Using the compactness of trajectories, one sees that there exists a subsequence $\{y_{k_\nu}\}$ converging uniformly to a certain admissible trajectory y. In particular,

$$y(t_0) = x^0 \text{ and } \varphi(y(T)) = \inf_{u\in L^\infty_U[t_0,T]} \varphi(x^{t_0,x^0,u}(T)).$$

Thus, for a certain $u_* \in L^\infty_U[t_0, T]$, we have $y = x^{t_0,x^0,u_*}$. We see that it is an optimal trajectory, and that u_* is an optimal control. This completes the proof of the existence theorem.

The only thing that we have to prove now is the compactness property. This requires a special tool from the theory of differential inclusions, which is called Filippov's lemma.

4.5 Fillipov's lemma and differential inclusions.

Definition. A multi-function Γ from $\mathbb{R}^m$ to $\mathbb{R}^n$ associates to every $y \in \mathbb{R}^m$ a set $\Gamma(y) \subset \mathbb{R}^n$ (possibly empty).

Definition. Let Γ be a multi-function from $\mathbb{R}^n$ to $\mathbb{R}^n$. We say that a function $y \in AC_{\mathbb{R}^n}[t_0, t_1]$ is a solution to the differential inclusion

$$y' \in \Gamma(y)$$

if for a.e. $t \in [t_0, t_1]$, $y'(t) \in \Gamma(y(t))$.

Example. With (CS) we associate the multifunction from $\mathbb{R}^n$ to $\mathbb{R}^n$

$$\text{that maps } x \in \mathbb{R}^n \text{ to } \mathbb{F}(x) = f(x, U) \subset \mathbb{R}^n.$$

Clearly, if $x(\cdot)$ is a solution to (CS), then $x(\cdot)$ is a solution of the differential inclusion

$$x' \in \mathbb{F}(x).$$

The converse to the last statement is also true, but is not obvious. It is given by the next result.

Filippov's lemma. *Let $x : [t_0, t_1] \to \mathbb{R}^n$ be a solution to the differential inclusion $x' \in \mathbb{F}(x)$. Then there exists a measurable $u : [t_0, t_1] \to U$ such that $x'(t) = f(x(t), u(t))$ for a.e. $t \in [t_0, t_1]$.*

The theory of set-valued analysis is a subject of the monographs [AC, AF].

Note that (H0) was not assumed. If we assume additionally (H0) then $u \in L^\infty_U[t_0, t_1]$ and $x(\cdot)$ is a solution to the differential equation $x'(t) = f(x(t), u(t))$.

4.6 A little bit of convex analysis.

Lemma. *Let $A, B \subset \mathbb{R}^n$ and*

$$A + B := \{a + b : a \in A, b \in B\}.$$

(i) If A and B are convex, then $A + B$ is convex,

(ii) If A is closed and B is compact, then $A + B$ is closed.

Remark. It is easy to give an example where A and B are closed, but $A + B$ is not closed.

Theorem (strong separation theorem)**.** *Let $S_1, S_2 \subset \mathbb{R}^n$ be convex and disjoint. Let S_1 be closed and let S_2 be compact. Then there exists $p \in \mathbb{R}^n$ and $\varepsilon > 0$ such that*

$$p \cdot x + \varepsilon \leq p \cdot y \text{ for all } x \in S_1,\ y \in S_2.$$

Here $p \cdot x$ is the scalar product.

Lemma. *Let $S \subset \mathbb{R}^n$ be closed and convex. Let $y \in L^1_{\mathbb{R}^n}[0, T]$ be such that $y(t) \in S$ a.e. and $v = \frac{1}{T}\int_0^T y(t)\mathrm{d}t$. Then $v \in S$.*

Proof. Suppose that $v \notin S$. Applying the strong separation theorem to $S_1 = S$ and $S_2 = \{v\}$, we see that there exists p and ε such that $p \cdot x + \varepsilon \leq p \cdot v$ for $x \in S$. Then $p \cdot v = \frac{1}{T}\int_0^T p \cdot v(t)\mathrm{d}t \leq p \cdot v - \varepsilon$, a contradiction. □

4.7 Proof of the compactness property.

Let $I = [t_0, t_1]$. Since $\{y_k\}$ is uniformly bounded, (UBf) implies that $\{y_k\}$ is uniformly Lipschitz continuous. Hence $\{y_k\}$ is uniformly equicontinuous. Thus, the Ascoli-Arzela theorem is applicable and gives $y_{k_\nu} \to y$ uniformly for a certain $y \in C_{\mathbb{R}^n}[t_0, t_1]$.

Let us recall the Ascoli-Arzela theorem and equicontinuity (see e.g. [RS1]).

Definition. A family $\{y_\alpha\}$ of functions on I is said to be uniformly equicontinuous if

$$\forall \epsilon > 0\ \exists \delta > 0 \text{ s.t. } |t - \widetilde{t}| < \delta \Rightarrow |y_\alpha(t) - y_\alpha(\widetilde{t})| < \epsilon \quad \forall y_\alpha.$$

A family $\{y_\alpha\}$ of functions is said to be equicontinuous if

$$\forall \epsilon > 0\ \forall t \in I \quad \exists \delta > 0 \text{ s.t. } |t - \widetilde{t}| < \delta \Rightarrow |y_\alpha(t) - y_\alpha(\widetilde{t})| < \epsilon \quad \forall y_\alpha.$$

Remark. An equicontinuous family of functions on I is uniformly equicontinuous (converse is obvious).

Theorem (Ascoli-Arzela)**.** *Let $\{y_\alpha\}$ be a family of uniformly bounded equicontinuous functions on I. Then some subsequence$\{y_{\alpha_k}\}$ converges uniformly on I.*

Without loss of generality we can redenote $\{y_{k_\nu}\}$ as $\{y_k\}$. So from now on $y_k(\cdot) \to y(\cdot) \in C_{\mathbb{R}^n}(I)$ uniformly. We only need to show that $y(\cdot)$ is a trajectory.

By Filippov's lemma, to show that $y(\cdot)$ is a trajectory of (CS) it is enough to prove that $y'(t) \in f(y(t), U)$ for a.a. t.

The uniform convergence implies that $y(\cdot) \in \mathrm{Lip}(I)$ (i.e., that y is uniformly Lipschitz continuous in I). So $y'(t)$ exists a.e..

Let $y[I] := \{y(t) : t \in I\}$. There exists $R > 0$ s.t.

$$y[I] \subset \overline{\bigcup_k y_k[I]} \subset B_R(0).$$

Let $M = const$ be such that $|f(x,u)| \leq M$ for all $x \in B_R(0)$, $u \in U$.

Let $\varepsilon > 0$ and $\mathbb{F}_\varepsilon(x) = f(x,U) + \overline{B_\epsilon(0)}$. Note that $\mathbb{F}_\varepsilon(x)$ is closed and convex due to the lemma above.

Let t be s.t. (such that) $y'(t)$ exists. From (H1), we have

$$\begin{aligned} |f(y_k(s),u_k(s)) - f(y(t),u_k(s))| &\leq K_1|y_k(s) - y(t)| \leq K_1(|y_k(s) - y_k(t)| + |y_k(t) - y(t)|) \\ &\leq K_1(M|s-t| + \|y_k - y\|_{L^\infty}). \end{aligned}$$

Hence for large enough k and small enough $|s-t|$, $f(y_k(s),u_k(s)) \in \mathbb{F}_\varepsilon(y(t))$. Then

$$\frac{y_k(t+h) - y_k(t)}{h} = h^{-1}\int_t^{t+h} f(y_k(s),u_k(s))\mathrm{d}s \in \mathbb{F}_\varepsilon(y(t))$$

for small h and large k. We have used the convexity and closedness of $\mathbb{F}_\varepsilon(x)$ and the lemma about an average.

Let now $k \to \infty$. We get $\frac{y(t+h)-y_k(t)}{h} \in \mathbb{F}_\varepsilon(y(t))$. Let $h \to 0$. Then we get $y'(t) \in \mathbb{F}_\varepsilon(y(t))$. Letting $\varepsilon \to 0$, we finally obtain $y'(t) \in f(y(t),U)$ and this is valid at every point t where $y'(t)$ exists. This completes the proof.

References for Section 4.

[AC] Aubin, J. P., Cellina, A. (2012). Differential inclusions: set-valued maps and viability theory (Vol. 264). Springer Science & Business Media.

[AF] Aubin, J. P., Frankowska, H. (2009). Set-valued analysis. Springer Science & Business Media.

[CS] Cannarsa, P., Sinestrari, C., Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004.

[RS1] Reed, M. and Simon, B., 1980. Methods of Modern Mathematical Physics (Revised ed.), Volume I: Functional Analysis.

5 Value function and dynamic programming principle.

We assume $f \in C_{\rm loc}(\mathbb{R}^n \times \mathbb{R}^m)$, (H0), (H1), and $\varphi \in C_{\rm loc}(\mathbb{R}^n)$. We DO NOT assume convexity of $f(x, U)$ and so do not know if optimal control exists.

Definition (value function for Mayer OCP)**.** The value function V maps $(t, y) \in [0, T] \times \mathbb{R}^n$ to

$$V(t, y) := \inf\{\varphi(x^{t,y,u}(T)) : u \in L^\infty_U[t, T]\}.$$

Note that here t plays the role of t_0 and $y \in \mathbb{R}^n$ the role of x^0 in (IS). The idea is that we play with the pair (t_0, x^0) of the initial time and the initial state (the (IS)-pair) and map the optimal cost for each such a pair.

So $V : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ and

$$V(T, y) = \varphi(y).$$

Theorem (Dynamical Programming Principle (DPP))**.**

(1) For any $s \in [t, T]$,

$$V(t, y) = \inf_{u \in L^\infty_U[t,s]} V\left(s, x^{t,y,u}(s)\right). \tag{DPP1}$$

(2) A control $u(\cdot) \in L^\infty_U[t, T]$ is optimal for the (IS)-pair (t, y) (i.e., for the initial condition $x(t) = y$) if and only if

$$V(t, y) = V(s, x^{t,y,u}(s)) \quad \forall s \in [t, T]. \tag{DPP2}$$

Proof. Step 1, "$\le$" in (DPP1). Let $\varepsilon > 0$ and $y_1(\cdot) = x^{s,y,u}(\cdot)$ for a certain $u \in L^\infty_U[t, s]$. Then there exists $v \in L^\infty_U[s, T]$ s.t.

$$\varphi(x^{s,y_1,v}(T)) \le V(s, y_1) + \varepsilon.$$

Let $w(\cdot)$ be the concatenation of $u(\cdot)$ and $v(\cdot)$, i.e., $w := u$ before s, and $w := v$ after s. Then $x^{t,y,w}(T) = x^{s,y_1,v}(T)$ and so

$$V(t, y) \le \varphi(x^{t,y,w}(T)) = \varphi(x^{s,y_1,v}(T)) \le V(s, y_1) + \varepsilon.$$

Letting $\varepsilon \to 0$, we get

$$V(t, y) \le V(s, y_1).$$

Since the control u is arbitrary, we get "$\le$" in (DPP1).

Step 2, "$\ge$" in (DPP1). Let a control $w(\cdot)$ be s.t.

$$V(t, y) \ge \varphi(x^{t,y,w}(T)) - \varepsilon.$$

Let u and v be restrictions of w to $[t, s)$ and $(s, T]$, resp.. Then

$$\inf_{\widetilde{u} \in L^\infty_U[t,s]} V\left(s, x^{t,y,\widetilde{u}}(s)\right) \le V(s, y_1) \le \varphi(x^{s,y_1,v}(T)) = \varphi(x^{t,y,w}(T)) \le V(t, y) + \varepsilon.$$

This gives "$\ge$" in (DPP1).

Step 3. To prove "if" in (2), we put $y_1 = x^{t,y,u}(s)$ for an optimal $u(\cdot) \in L^\infty_U[t, T]$. Then

$$V(t, y) = \varphi(x^{t,y,u}(T)) = \varphi(x^{s,y_1,u}(T)) = V(s, y_1).$$

To prove "only if", we plug $s = T$ in (DPP2) and get $V(t, y) = \varphi(x^{t,y,u}(T))$. So x and u are minimizers. □

Corollary. *Let $x(\cdot)$ be the trajectory corresponding to a control $u(\cdot)$. Then:*

(1)

$$V(t,y) \leq V(s,x(s)), \quad s \in [t,T]$$

(2) The equality

$$V(t,y) = V(s,x(s)) \quad \textit{holds for all } s \in [t,T] \tag{**}$$

if and only if u and x are optimal.

References for Section 5.

[CS] Cannarsa, P., Sinestrari, C., Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004.

6 Hamilton-Jacobi-Bellman (HJB) equation and viscosity solutions.

6.1 HJB equation for Mayer's Problem.

We assume (H0), (H1), $f \in C_{\text{loc}}(\mathbb{R}^n \times \mathbb{R}^m)$, $\varphi \in C_{\text{loc}}(\mathbb{R}^n)$, and $J(u) = J(u, x^0, t_0) = \varphi(x(T))$. We DO NOT assume convexity of $f(x, U)$ and so do not know if optimal control exists.

Definition. The Hamiltonian function for MOCP is $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\mathcal{H}(x, p) := \max_{u \in U}(-p \cdot f(x, u)).$$

The Hamilton-Jacobi-Bellman equation associated with MOCP is

$$-\partial_t v(t, y) + \mathcal{H}(y, \partial_y v(t, y)) = 0, \tag{HJB}$$

where $\partial_t v = \partial v/\partial t$, and $\partial_y v = (\partial v/\partial y_1; \partial v/\partial y_2; \ldots; \partial v/\partial y_n)$. (HJB) is equipped with the terminal value condition

$$v(T, y) = \varphi(y). \tag{TVC}$$

Recall that the value function V satisfies (TVC) by definition.

Theorem. *Assume that $\varphi \in Lip_{\text{loc}}(\mathbb{R}^n)$. Then $V \in Lip_{\text{loc}}([0, T] \times \mathbb{R}^n)$ and the value function V is a unique viscosity solution to the problem (HJB), (TVC).*

The proof is postponed for several lectures. Under additional assumptions it is easy to prove the classical pointwise version.

Theorem. *Assume that V is differentiable at a point $(t, y) \in [0, T] \times \mathbb{R}^n$ and there exists an optimal control strategy $u_*(\cdot)$ for the (IS)-pair (t, y) with the property that $u_*(\cdot)$ has a right limit $u_*(t + 0)$. Then V satisfies (HJB) at (t, y) and*

$$-\partial_t V(t, y) - \partial_y V(t, y) \cdot f(y, u_*(t + 0)) = 0. \tag{*}$$

Proof. Let $w(\cdot) \in L^\infty_U[t, T]$ and let $x(s) = x^{t,y,w}(s)$. Then $V(t, y) \le V(t + h, x(t + h))$ due to (DPP1). So, as $h \to 0$,

$$0 \le V(t + h, x(t + h)) - V(t, y) = \partial_t V(t, y)h + \partial_x V(y) \cdot (x(t + h) - x(t)) + o(h).$$

Let w be s.t. $x'(t)$ exists, e.g., we can take for $s \in [t, t + \delta]$, $w(s) = const = u \in U$. Then $x'(t) = f(y, u)$ and we have

$$0 \le V(t + h, x(t + h)) - V(t, y) = h\partial_t V(t, y) + h\partial_x V(y) \cdot x'(t) + o(h).$$

and so

$$0 \le \partial_t V(t, y) + \min_{u \in U}(\partial_y V \cdot f(y, u)).$$

Taking $w = u_*$ we get $x'(t) = f(y, u_*(t + 0))$.

On the other hand, from (DPP2) we have

$$0 = V(t + h, x(t + h)) - V(t, y) = \partial_t V(t, y)h + h\partial_x V(y) \cdot f(y, u_*(t + 0)) + o(h).$$

This gives (*) and (HJB) at (t, y). □

6.2 Viscosity solutions.

Let $A \subset \mathbb{R}^n$ be open.

6.2.1 Generalized differentials.

Let $v : A \to \mathbb{R}$.

Definition. For $x \in A$,

$$D^- v(x) = \left\{ p \in \mathbb{R}^n \ : \ \liminf_{y \to x} \frac{v(y) - v(x) - \langle p, y - x \rangle}{|y - x|} \geq 0 \right\}$$
$$D^+ v(x) = \left\{ p \in \mathbb{R}^n \ : \ \limsup_{y \to x} \frac{v(y) - v(x) - \langle p, y - x \rangle}{|y - x|} \leq 0 \right\}$$

are, resp., superdifferential and subdifferential of v at x ($\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^n$).

Lemma.

(1) $D^-(-v)(x) = -D^+ v(x)$,

(2) $D^{\pm} v(x)$ *are closed convex sets (possibly empty)*

(3) $D^{\pm} v(x)$ *are both nonempty if and only if* v *is differentiable at* x*; in this case*

$$D^+ v(x) = D^- v(x) = \{ Dv(x) \}$$

Here $Dv(x) = \partial_x v(x)$ is the gradient.

Proof. (1) and (2) are obvious from the definition. (3) If v is differentiable at x, then $D^+ v(x) = D^- v(x) = \{ Du(x) \}$ easily follows from definition of differentiability. For the proof of the part "if" of (3), see [CS]. □

Example.

(1) Let $A = \mathbb{R}$ and $v(x) = |x|$. Then $D^+ v(0) = \varnothing$, $D^- v(0) = [-1, 1]$.

(2) Let $A = \mathbb{R}$ and $v(x) = |x|^{1/2}$. Then $D^+ v(0) = \varnothing$, $D^- v(0) = \mathbb{R}$.

(3) Let $A = \mathbb{R}^2$ and $v(x) = |x_1| - |x_2|$. Then $D^{\pm} v(0) = \varnothing$.

Definition. We say that $\psi(\cdot)$ touches v from above (from below) at $x_0 \in A$ if $v(x_0) = \psi(x_0)$ and for all x in some small open ball $B_r(x_0)$

$$v(x) \leq \psi(x) \quad \text{(resp., } v(x) \geq \psi(x)\text{)}.$$

Lemma 6.1. *Let* $v \in C_{\mathrm{loc}}(A)$, $p \in \mathbb{R}^n$, *and* $x \in A$. *Then the following statements are equivalent:*

(1) $p \in D^+ v(x)$ *(resp.,* $p \in D^- v(x)$*);*

(2) $p = D\psi(x)$ *for some* $\psi \in C^1(A)$ *touching* v *from above (resp., below) at* x.

(3) $p = D\psi(x)$ *for some* $\psi \in C^1(A)$ *s.t.* $v - \psi$ *attains a local maximum (resp., minimum) at* x.

For the complete proofs of the lemma, see [CS, Section 3].

6.3 Definition of viscosity solution

Let $A \subset \mathbb{R}^n$ be open. Let $F(\cdot) \in C_{\mathrm{loc}}(A \times \mathbb{R} \times \mathbb{R}^n)$. Consider the equation

$$F(x, v, Dv) = 0, \qquad x \in A \subset \mathbb{R}^n. \tag{GenEq}$$

The HJB equation can be written in this form if we recast t as a component, say x_0 of the space variable $x = (x_0, x_1, \ldots, x_{n-1})$.

Definition. A function $v \in C_{\mathrm{loc}}(A)$ is called a viscosity subsolution to (GenEq) if

$$F(x, v(x), p) \le 0 \qquad \text{for all } p \in D^+v(x) \text{ and for all } x \in A. \tag{SubS}$$

A function $v \in C_{\mathrm{loc}}(A)$ is called a viscosity supersolution to (GenEq) if

$$F(x, v(x), p) \ge 0 \qquad \text{for all } p \in D^-v(x) \text{ and for all } x \in A. \tag{SupS}$$

If v satisfies both (SubS) and (SupS), it is called a viscosity solution in A.

Remark. If v is differentiable at x, then the combination of (SubS) and (SupS) at x is equivalent to

$$F(x, v(x), Dv(x)) = 0.$$

So if (GenEq) possesses a classical solution v in the sense that v differentiable at every $x \in A$ and (GenEq) holds, then v is a viscosity solution in A.

6.4 The HJB equation for Bolza problem with a fixed terminal time

For $(t, y) \in [0, T] \times \mathbb{R}^n$, let us consider the cost functional

$$J_{t,y}(u) = \int_t^T \ell(x(s), u(s))\mathrm{d}s + \varphi(x(T)),$$

where $x(\cdot) = x^{t,y,u}(\cdot)$.

Bolza optimal control problem with fixed T (BOCP). Minimize $J_{t,y}$ over $u \in L^\infty_U[t, T]$.

The case $\ell = 0$ leads to Mayer OCP.

It is possible to transform BOCP to MOCP. Let $X = (x_0, x)^\top = (x_0, x_1, \ldots, x_n)^\top$. Let us define a new control system $X' = \widetilde{f}(X, u)$, where $\widetilde{f} : \mathbb{R} \times \mathbb{R}^n \times U \to \mathbb{R} \times \mathbb{R}^n$ is defined by

$$\widetilde{f}(X, u) = (\ell(x, u); f(x, u)) = \begin{pmatrix} \ell(x, u) \\ f_1(x, u) \\ \cdots \\ f_n(x, u) \end{pmatrix}.$$

Let us define a new terminal cost function $\widetilde{\varphi} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$,

$$\widetilde{\varphi}(x_0, x) = \widetilde{\varphi}(X(T)) := x_0 + \varphi(x).$$

Since $x_0(s) = \int_t^s \ell(x(s), u(s))$, we see that

$$J_{t,y}(u) = \widetilde{\varphi}(x_0(T), x(T)) = \widetilde{\varphi}(X(T)).$$

Thus, we have formally rewritten the Bolza problem as the Mayer problem.

To make this reduction rigorous, one have to impose some regularity assumptions on ℓ. The first assumption is standard $\ell \in C_{\rm loc}(\mathbb{R}^n \times \mathbb{R}^m)$ (so that $\widetilde{f} \in C_{\rm loc}$).

The assumption of Lipschitz continuity in x uniform with respect to U can be relaxed for ℓ and replaced by the assumption

(L1) for any $R > 0$ there exists γ_R s.t. $|\ell(x,u)| - \ell(\widetilde{x},u)| \le \gamma_R |x - \widetilde{x}|$ for all $x, \widetilde{x} \in B_R(0)$ and $u \in U$.

Indeed, under assumption (H0) we had a lemma that states that for any $r > 0$ there exists R such that $|x^{t,y,u}(s)| < R$ whenever $y \in B_r(0)$ and $u \in L^\infty_U$. So for $t, s \in [0,T]$ we in any case use (L1) only on bounded sets of x.

The value function for BOCP is defined by

$$V(t,y) = \inf_{u \in L^\infty_U[t,T]} J_{t,y}(u).$$

The Hamiltonian function associated with BOCP is

$$\mathcal{H}(x,p) := \max_{u \in U} \left(-p \cdot f(x,u) - \ell(x,u)\right),$$

that is, the associated HJB equation $\partial_t v(t,y) = \mathcal{H}(y, \partial_y v(t,y))$ becomes

$$-\partial_t v(t,y) + \max_{u \in U} \Big(-\langle \partial_y v(t,y), f(x,u) \rangle_{\mathbb{R}^n} - \ell(x,u) \Big) = 0. \tag{HJB}$$

Theorem. *Assume (H0), (H1), $\ell(\cdot,\cdot) \in C_{\rm loc}(\mathbb{R}^n \times \mathbb{R}^m)$, and (L1). Assume that V is differentiable at a point (t,y) and there exists an optimal control strategy $u_*(\cdot)$ for (IS) $x(t) = y$ with the property that $u_*(\cdot)$ has a right limit $u_*(t+0)$. Then V satisfies (HJB) at (t,y) and*

$$-\partial_t V(t,y) - \partial_y V(t,y) \cdot f(y, u_*(t+0)) - \ell(y, u_*(t+0)) = 0.$$

The proof is essentially the same as for the Mayer problem.

References for Section 6.

[CS] Cannarsa, P., Sinestrari, C., Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004.

[E] L.C. Evans, Lecture notes of the course "An Introduction to Mathematical Optimal Control Theory", https://math.berkeley.edu/ evans/control.course.pdf

[FR] Fleming, W. H., Rishel, R. W. (2012). Deterministic and stochastic optimal control (Vol. 1). Springer Science & Business Media.

7 Stochastic differential equations.

This part is intended to give some impression of stochastic optimal control considering the simplest stochastic differential equation (SDE)

$$\begin{cases} X'(s,\omega) = f(X(s,\omega), u(t,\omega)) + \gamma\xi(s,\omega), \quad s \in [t,T] \\ X(t,\omega) = y \end{cases}$$

with a white noise $\xi(t)$, and to write the corresponding HJB equation following mainly [E].

7.1 Stochastic processes.

A solution $x : [t,T] \times \Omega \to \mathbb{R}^n$ to SDE is a random process, that is, roughly speaking, a function depending on the time t and on the sample point $\omega \in \Omega$, where $(\Omega, \Sigma, \mathbb{P})$ is a probability space.

Let $\mathcal{B}$ be σ-algebra of Borel subsets of $\mathbb{R}^n$. A mapping $X : \Omega \to \mathbb{R}^n$ is an n-dim. random variable (r.v.) if $X^{-1}(S) \in \Sigma$ for any $S \in \mathcal{B}$. If this is the case $\Sigma(X) := \{X^{-1}(S) : S \in \mathcal{B}\}$ is the σ-algebra generated by X.

Example. Let X be a 1-D r.v.. If the probablity $\mathbb{P}(X < x)$ has the form $\int_{-\infty}^{x} f(y)\mathrm{d}y$ with the density function

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{|x-\mu|}{2\sigma^2}}, \quad y \in \mathbb{R},$$

then it is said that X has a Gaussian (or normal) distribution with mean (math. expectation) μ and variance σ^2. We will write $X \in N(\mu, \sigma^2)$.

A collection $\{X(t) = X(t,\omega) : t \geq 0\}$ of random variables is called a stochastic process (s.p.). For any fixed sample point $\omega \in \Omega$, the mapping $t \mapsto X(t,\omega)$ is called a sample path of the s.p. $X(t)$, $t > 0$, corresponding to ω.

Definition. An $\mathbb{R}$-valued stochastic process $W(t)$, $t \geq 0$, is called a 1-dim. Wiener process (or 1-dim. Brownian motion) if

(i) $W(0) = 0$ a.s.,

(ii) $W(t) - W(s) \in N(t-s)$ for all $0 \leq s < t$,

(iii) for any finite partition $0 < t_1 < t_2 < \cdots < t_n$ of $[0, +\infty)$, the r.v. $W(t_1)$, $W(t_2) - W(t_1)$, $\ldots$, $W(t_n) - W(t_{n-1})$ are independent.

An accessible explanations of the theorem of existence of a Wiener process can be found in [E-SDE].

Something like a "definition" of a 1-dim. white noise $\xi(t)$. A Wiener process is "a unique solution" to the following SDE

$$\begin{cases} X'(t) = \xi(t), \quad t \in [0, +\infty) \\ X(0) = 0 \text{ a.s.} \end{cases}$$

Definition. A s.p $\widetilde{X}(\cdot)$ is called a version of a s.p. $X(\cdot)$ if $\mathbb{P}(X(t) = \widetilde{X}(t)) = 1$ for all $t \geq 0$.

By $C^{0,\alpha}[0,T]$ we denote the space of uniformly Hölder continuous on $[0,T]$ functions with exponent $\alpha \in (0,1)$.

Theorem. *Let $T > 0$. Let $W(\cdot)$ be a Wiener process. Then:*

(i) There exists a version $\widetilde{W}(\cdot)$ such that $X(\cdot,\omega) \in C^{0,\alpha}[0,T]$ for each $\alpha \in (0,1/2)$ with probability 1 (i.e., for a.a. ω w.r.t. the probability measure $\mathbb{P}$).

(ii) For each $\alpha \in (1/2,1)$, with probability 1, $W(\cdot,\omega)$ is nowhere Hölder continuous with exponent α (pointwise in t).

(iii) Almost surely, $W(\cdot,\omega)$ is of infinite variation on each interval $[t_1,t_2] \subset [0,T]$ ($t_1 < t_2$) and the sample path $t \to W(t,\omega)$ is nowhere differentiable.

Definition. Let $\{\Sigma_j\}_{j\in J}$ be a family of σ-algebras $\Sigma_j \subset \Sigma$.

(i) The σ-algebras Σ_j are said to be independent if for any $j_1, \dots j_k$, and $A_{j_i} \in \Sigma_{j_i}$, the collection of events $\{A_{j_i}\}_{i=1}^k$ is independent.

(ii) By $\Sigma(\Sigma_j, j \in J)$ we denote the smallest σ-algebra $\widetilde{\Sigma} \subset \Sigma$ that contain all Σ_j.

(iii) $\mathcal{W}^+(t) := \Sigma(W(s) - W(t) : s \geq t)$ is the future of the Wiener process beyond the time t.

Definition. An $\mathbb{R}^n$-valued stochastic process $W(\cdot)$ is called a n-dim. Wiener process (or n-dim. Brownian motion) if

(i) each of coordinates $W_k(\cdot)$, $k = 1,\dots,n$, is a 1-dim. Wiener process,

(ii) $\mathcal{W}_k := \Sigma\left(\Sigma(W_k(t),\ t \geq 0)\right)$, $k = 1,\dots,n$, are independent.

7.2 A particular case of Itô's chain rule.

Let $W(\cdot)$ be n-D Brownian motion. Let X^0 be a r.v..

Consider

$$\begin{cases} X'(t) = f(t,X(t)) + \gamma\xi(t), & t \in [t_0,T] \\ X(t_0) = X^0 \end{cases}$$

In the theory of SDE this is usually written as

$$\begin{cases} \mathrm{d}X(t) = f(t,X(t))\mathrm{d}t + \gamma \mathrm{d}W(t), & s \in [t_0,T] \\ X(t_0) = X^0 \end{cases} \tag{SDE}$$

We want to specify conditions on f, X^0, and a s.p. $X(s)$ so that it makes sense to say that

$$X(t) = X^0 + \int_{t_0}^t f(\tau,X(\tau))\mathrm{d}\tau + \gamma(W(t) - W(t_0)).$$

is a solution to (SDE). These conditions also should be imposed in such a way that they will work for the case with stochastic control $u(t,\omega)$.

First, note that we need a definition of the integral not only w.r.t. time $\int \cdot\, \mathrm{d}t$, but also w.r.t. $\mathrm{d}W$. Indeed, to derive HJB equation we need to plug the s.p. $X(s)$ into the value function $V(t,x)$ and in some sense to differentiate the result $V(t,X(t))$. This can be done with the use of Itô's formula. Let us consider a particular case of it, first, on a non-rigorous level.

Suppose that a 1-dim. s.p. $X(\cdot)$ satisfies $\mathrm{d}X = F\mathrm{d}t + \gamma \mathrm{d}W$ for $t \in [t_0, T]$ in the sense that

$$X(t) = X(t_0) + \int_{t_0}^{t} F(s)\mathrm{d}s + \gamma(W(t) - W(t_0)), \qquad t \geq t_0,$$

where $F(\cdot)$ is a s.p. with a good enough properties, which we have not specified yet.

Assume that $v(t,x)$, $v : [0,T] \times \mathbb{R} \to \mathbb{R}$, is continuous together with $\partial v/\partial t$, $\partial v/\partial x$, $\partial^2 v/\partial x^2$. Consider the s.p. $Y(t) := v(t, X(t))$. Then

$$\mathrm{d}Y = \frac{\partial v}{\partial t}\mathrm{d}t + \frac{\partial v}{\partial x}\mathrm{d}X + \frac{\gamma^2}{2}\frac{\partial^2 v}{\partial x^2}\mathrm{d}t = \left(\frac{\partial v}{\partial t}\mathrm{d}t + \frac{\partial v}{\partial x}F + \frac{\gamma^2}{2}\frac{\partial^2 v}{\partial x^2}\right)\mathrm{d}t + \gamma\frac{\partial v}{\partial x}\mathrm{d}W,$$

where $\frac{\partial v}{\partial t}$ means $\frac{\partial v}{\partial t}(t, X(t))$, etc.. The equality have to be understood in the intergal sense.

Note that the last term is

$$\gamma\frac{\partial v}{\partial x}(t, X(t))\mathrm{d}W.$$

So one needs a definition of

$$\int_{t_0}^{t} Z(t)\mathrm{d}W(t),$$

where $Z(t)$ is a s.p.

Rough explanation of Itô's chain rule. Let us write for small h,

$$\delta t = h, \quad \delta X = X(t+h) - X(t), \quad \delta W = W(t+h) - W(t) \text{ etc.}$$

Recall that $W(\cdot)$ is continuous a.s., so "$\delta W = o(1)$". Writing the Taylor formula for $Y(t+h)$ and using the "rule"

$$(\delta W)^2 \approx \delta t,$$

we get

$$\begin{aligned}
\delta Y &= \frac{\partial v}{\partial t}\delta t + \frac{\partial v}{\partial x}\delta X + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}(\delta X)^2 + o(\delta t + (\delta X)^2) \\
&= \frac{\partial v}{\partial t}\delta t + \frac{\partial v}{\partial x}(F\delta t + \gamma\delta W + o(\delta t)) + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}(F\delta t + \gamma\delta W + o(\delta t))^2 + o(\delta t + (\delta X)^2) \approx \\
&\approx \frac{\partial v}{\partial t}\delta t + \frac{\partial v}{\partial x}(F\delta t + \gamma\delta W) + \frac{\gamma^2}{2}\frac{\partial^2 v}{\partial x^2}\delta t + o(\delta t).
\end{aligned}$$

To give a less rough explanation, let us consider a rigorous version of the rule $(\delta W)^2 \approx \delta t$.

Lemma. *Let $W(\cdot)$ be a 1-dim. Brownian motion and let $[a,b] \subset [0,+\infty)$. Let P be a partition $a = t_0 < t_1 < \cdots < t_m = b$ of $[a,b]$ with the size $|P| := \max_k |t_{k+1} - t_k|$. Let*

$$Q = \sum_{k=0}^{m-1} [W(t_{k+1}) - W(t_k)]^2.$$

Then $Q \to b - a$ as $|P| \to 0$ in the sense of $L^2(\Omega)$, i.e.,

$$\mathbb{E}\left(|Q - (b-a)|^2\right) \to 0.$$

Proof.

$$Q-(b-a)=\sum_{k=0}^{m-1}\left([W(t_{k+1})-W(t_k)]^2-(t_{k+1}-t_k)\right),$$

$$\mathbb{E}(|Q-(b-a)|^2)=$$

$$=\sum_{k=0}^{m-1}\sum_{j=0}^{m-1}\mathbb{E}\left[\left([W(t_{k+1})-W(t_k)]^2-(t_{k+1}-t_k)\right)\left([W(t_{j+1})-W(t_j)]^2-(t_{j+1}-t_j)\right)\right]$$

$$=\sum_{k=0}^{m-1}\mathbb{E}\left([W(t_{k+1})-W(t_k)]^2-(t_{k+1}-t_k)\right)^2=\sum_{k=0}^{m-1}\mathbb{E}\left[(Z_k^2-1)^2(t_{k+1}-t_k)^2\right],$$

where $Z_k=\frac{W(t_{k+1})-W(t_k)}{(t_{k+1}-t_k)^{1/2}}\in N(0,1)$. Thus,

$$E(|Q-(a-b)|^2)\le C\sum_{k=0}^{m-1}(t_{k+1}-t_k)^2\le C|P|(b-a)$$

and the right and the left sides of this formula go to 0 as $|P|\to 0$. □

We see that the main cancellation happened because for $j\ge k+1$, $W(t_{j+1})-W(t_j)$ (future) are independent of $W(t_{k+1})-W(t_k)$ (past).

Let us now try to define the simplest integral of the form $\int_0^T Z(t)\mathrm{d}X(t)$. Namely, let us take a partition P of $[0,T]$ and consider $\int_0^T W(t)\mathrm{d}W(t)$ taking a specific form of Riemann sum formally associated with $\int_0^T W(t)\mathrm{d}W(t)$:

$$R_P=\sum_{k=0}^{m-1}W(t_k)[W(t_{k+1})-W(t_k)].$$

Sums of this type correspond to *the Itô integral*.

Lemma. *$R_P\to W(T)^2/2-T/2$ as $|P|\to 0$ in the $L^2(\Omega)$-sense.*

Proof. Using summation by parts, we see that

$$R_P=\frac{W(T)^2}{2}-\frac{1}{2}\sum_{k=0}^{m-1}[W(t_{k+1})-W(t_k)]^2\to\frac{W(T)^2}{2}-\frac{T}{2}. \tag{4}$$

as $|P|\to 0$ due to the previous lemma. □

So it is naturally to assume that

$$\int_0^T W\mathrm{d}W=W(T)^2/2-T/2.$$

7.3 Filtration and progressive measurability*.

We have seen that the main cancellation happens because for $j \geq k+1$, $W(t_{j+1}) - W(t_j)$ are independent of $W(t_{k+1}) - W(t_k)$ (the future is independent of the past). It is important for building *the Itô integral* and for the definition of solutions $X(s)$ of SDE that the property of independence of the future $\mathcal{W}^+(t) := \Sigma(W(s) - W(t), t \leq s)$ of the Brownian motion is preserved.

The notion of filtration is devised to ensure this. Assume that the r.v. X^0 is independent of $\mathcal{W}^+(0)$. We will use the filtration

$$\mathbb{F} = \{\mathbb{F}_t\}_{t \geq 0}, \text{ where } \mathbb{F}_t := \Sigma(X^0, W(s), s \in [0,t]).$$

(Note that $\mathbb{F}_t$ and $\mathcal{W}^+(t)$ are independent by the definition of the Brownian motion.)

We will assume that a solution $X(\cdot)$ of SDE is adapted to $\mathbb{F}$, i.e., $\Sigma(X(t)) \subset \mathbb{F}_t$.

The formula

$$X(s) = X^0 + \int_t^s f(\tau, X(\tau))\mathrm{d}\tau + \gamma(W(s) - W(t)) \qquad (*)$$

assumes that the integral $\int_t^s f(X(\tau), \tau)\mathrm{d}\tau$ of the s.p. $f(X(\tau), \tau)$ is a r.v. $X(s) - X^0 - \gamma(W(t) - W(s))$.

A technical peculiarity of stochastic integration is that an integral of a s.p. is not necessarily a r.v. if some additional conditions are not imposed.

Example. Let A be a non-measurable w.r.t. Lebesgue measure subset of $[0,1]$. Let us consider a deterministic function

$$f(s) = 1 \text{ for } s \in A, \qquad f(s) = -1 \text{ for } s \in [0, +\infty) \setminus A.$$

Let us define now a s.p. $X(s, \omega) := f(s)$ for all ω and $s \geq 0$. Then:

(1) $X(\cdot)$ is adapted to $\mathbb{F}$,

(2) there exist expectations

$$\mathbb{E}\int_0^t |X(s)|\mathrm{d}s = \mathbb{E}\int_0^t |X(s)|^2\mathrm{d}s = t.$$

(3) However, $\int_0^t X(s, \omega)\mathrm{d}s$ does not exist for any $\omega \in \Omega$, and so we cannot associate a certain r.v. with the integral $\int_0^t X(s)\mathrm{d}s$.

To handle this difficulty, one have to impose some assumption of measurability w.r.t. time t and to agree this with the filtration $\mathbb{F}$. This leads to a definition of progressive measurability.

Definition. A s.p. $X(\cdot)$ is called progressively measurable if for each $T \geq 0$ the function $X(t, \omega)$ on $[0,T] \times \Omega$ is measurable w.r.t. the minimal σ-algebra $\mathcal{B}_{[0,T]} \otimes \mathbb{F}_T$ generated by the sets $S \times A$, $S \in \mathcal{B}_{[0,T]}$, $A \in \mathbb{F}_T$, where $\mathcal{B}_{[0,T]}$ is the σ-algebra of Borel subsets of $[0,T]$.

There is another way to avoid the technical difficulty connected with the last example. We follow [GS].

Definition. Denote by $\mathcal{D}^n$ the space of n-dimensional functions $f(t)$, $t \in [0,T]$, such that:

(1) for each $t \in [0,T)$ there exists $f(t+0)$ and $f(t+0) = f(t)$,

(2) there exists $f(t-0)$ for each $t \in (0,T]$.

Such functions sometimes are called 'cadlag' functions.

When $\mathcal{D}^n$ is equipped with a special good metric, which makes it a separable metric space, $\mathcal{D}^n$ is called the Skorokhod space.

Definition. Denote by Φ, the class of s.p. $X(\cdot)$ such that $X(\cdot,\omega) \in \mathcal{D}^n$ for all ω and $X(\cdot)$ is adapted to $\mathbb{F}$.

If $X(\cdot) \in \Phi$, then $X(\cdot)$ is progressively measurable. That is why we will not impose assumptions of progressive measurability.

7.4 Solutions to SDE*.

Definition. A s.p. $X(t)$, $t \in [0,T]$, is called a solution to (SDE) on $[0,T]$ if the following conditions are satisfied:

(1) X is adapted to $\mathbb{F}$,

(2) $X(\cdot,\omega) \in \mathcal{D}^n$ with probability 1,

(3) for all t the formula (*) holds with probability 1.

To give existence and uniqueness theorem we impose several conditions on the function f following [GS] and [O].

Theorem (existence and uniqueness for SDE)**.** *Assume that the r.v. X^0 is independent of $\mathcal{W}^+(0)$ and s.t. $\mathbb{E}[|X^0|^2] < \infty$. Assume that $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ is measurable, satisfy the following local Lipschitz condition in x*

$$\forall R > 0 \ \exists K_R = const \ s.t. \ |f(t,x) - f(t,\widetilde{x})| \le K_R|x - \widetilde{x}| \quad \forall x, \widetilde{x} \in B_R(0) \qquad (\mathrm{H1}_{\mathrm{loc}})$$

and is linearly bounded, i.e.,

$$|f(t,x)| \le C(1+|x|) \qquad \forall x, t.$$

Then (SDE) has a unique solution $X(\cdot)$ in the space Φ.

Remark. The class Φ is actually not the appropriate choice for the above existence and uniqueness theorem. It was devised for a much wider class of SDE (see [GS, GS-C]).

References to Section 7.1.

[GS] Gihman, I.I. , Skorohod, A.V. (1979). Stochastic differential equations. Springer, New York, NY.

[GS-C] Gihman, I. I., Skorohod, A. V. (2012). Controlled stochastic processes. Springer Science & Business Media.

[E] L.C. Evans, Lecture notes of the course "An Introduction to Mathematical Optimal Control Theory", https://math.berkeley.edu/ evans/control.course.pdf

[E-SDE] L.C. Evans, Lecture notes of the course "An introduction to stochastic differential equations"

[O] Øksendal, B. Stochastic differential equations. Springer, 2003.

8 Stochastic optimal control problem.*

8.1 Various classes of admissible stochastic controls.

Consider now a controlled SDE

$$\begin{cases} \mathrm{d}X(t) = f(X(t), u(t))\mathrm{d}t + \gamma \mathrm{d}W(t) \\ X(0) = y \end{cases},$$

where $y \in \mathbb{R}^n$.

There are several classes of admissible controls $u(\cdot)$. The class $L_U^\infty[0,T]$ of deterministic admissible controls is rarely used in applications.

Let us consider several classes of s.p. $u(t,\omega)$ that can serve as classes of admissible controls.

Definition. The class $\widetilde{\mathbb{A}}$ of generalized controls consists of s.p. $u(t,\omega)$ taking values in U and such that u is progressively measurable.

Another class $\mathbb{A}_{feedback}$ of controls are feedback controls that are defined via functionals $u(t, x(\cdot))$ that are "good enough" and *non-anticipative* w.r.t. the second variable $x(\cdot) \in \mathcal{D}^n$. A functional $u(t,\cdot)$ is called non-anticipative if it maps a deterministic function $x(\cdot) \in \mathcal{D}^n$ to U such that for all t it follows from $x(s) = \widetilde{x}(s)$, $s \in [0,t]$, that $u(t, x(\cdot)) = u(t, \widetilde{x}(\cdot))$.

The control then take the form $u(t, X(\cdot))$, and the controlled SDE is

$$\begin{cases} \mathrm{d}X(t) = f(X(t), u(t, X(\cdot)))\mathrm{d}t + \gamma \mathrm{d}W(t) \\ X(0) = y \end{cases}.$$

There is an important subclass of feedback controls formed by *Markov* controls. These controls are defined by ''good enough" feedback functions $u : [0,T] \times \mathbb{R}^n \to U$. The control have the form $u(t, X(t))$ and the controlled SDE is

$$\begin{cases} \mathrm{d}X(t) = f\left(X(t), u(t, X(t))\right)\mathrm{d}t + \gamma \mathrm{d}W(t) \\ X(0) = y \end{cases}.$$

That is, the control $u(t, X(t))$ at a time t observe only the value $X(t)$ at this moment, and does not take into account the prehistory $X(s)$, $s \in [0,t)$.

In each case SDE requires some generalization. For example, for a fixed generalized control the expression $f(X(t), u(t))$ is not a function of type $F(X(t), t)$, but has the type $F(X(t), t, \omega)$. That is f itself depends on ω and so is random.

In the case of generalized controls the notion of solution to SDE and the existence and uniqueness theorem given above can be adapted to the case of $f = f(X(t), t, \omega)$ almost without changes [GS]. In other cases more care is needed.

8.2 Value functions and HJB equations for stochastic control.

Let $\mathbb{A}$ be a certain reasonable class of admissible controls $u(t,\omega)$.

Consider the controlled SDE

$$\begin{cases} \mathrm{d}X(s) = f(X(s), u(s, X(\cdot)))\mathrm{d}t + \gamma \mathrm{d}W(s), \quad s \in [t,T] \\ X(t) = y \end{cases}. \tag{cSDE}$$

Definition. Let $X(s) = X^{t,y,u}$ be the solution to (cSDE) corresponding to a control u. Then the functional $J_{t,y} : \mathbb{A} \to \mathbb{R}$ defined by

$$J_{t,y} := \mathbb{E}\left(\int_t^T \ell(X(s), u(s))\mathrm{d}s + \varphi(X(T))\right), \text{where } \mathbb{E}(\cdot) \text{ is the mathematical expectation,}$$

is called the expected cost functional in Bolza form.

The value function $V : [0, T] \times \mathbb{R}^n$ is defined in the way similar to that we had before,

$$V(t, y) = \inf_{u \in \mathbb{A}[t,T]} J_{t,y}(u)$$

The n-dim. Itô's chain rule. Let $X(\cdot)$ and $F(\cdot)$ be progressively measurable n-dim. stochastic processes such that

$$F \in L^2([0, T] \times \Omega), \quad \text{and} \quad \mathrm{d}X = F\mathrm{d}t + \gamma \mathrm{d}W.$$

Assume that $v(t, x)$, $v : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$, has "good enough differetiability properties" and consider the n-dim. s.p.

$$Y(t) := v(t, X(t)).$$

Then

$$\mathrm{d}Y = \frac{\partial v}{\partial t}\mathrm{d}t + \langle \frac{\partial v}{\partial x}, F\mathrm{d}t + \gamma \mathrm{d}W\rangle_{\mathbb{R}^n} + \frac{\gamma^2}{2}\sum_{j=1}^n \frac{\partial^2 v}{\partial x_j^2}\mathrm{d}t. \qquad (*)$$

The HJB equation corresponding the controlled stochastic equation (cSDE) is

$$-\partial_t V(t, y) - \frac{\gamma^2}{2}\Delta_y V(t, y) + \max_{u \in U}(-\partial_x V(t, y) \cdot f(y, u) - \ell(y, u)) = 0, \qquad \text{(sHJB)}$$

$$V(t, y) = \varphi(y), \qquad \text{(TV)}$$

where $\Delta_y V = \sum_{j=1}^n \frac{\partial^2 V}{\partial y_j^2}$.

The additional term $\frac{\gamma^2}{2}\Delta_y V$ comes from the n-dim Itô's chain rule. The rigorous derivation require more tools, for the case of Markov controls see [O, Section 11.2].

If coefficients of the equation and U are good enough then such equation have solutions with better properties. The uniqueness results can be established without notion of viscosity solution.

For a particular value of the parameter $\gamma \in (0, +\infty)$, let us denote the corresponding solution by V_γ. If, passing to the limit as $\gamma \to 0$, we obtain a limit $\lim_{\gamma \to 0} V_\gamma = V$ [CL83, FR] then, under certain assumptions, V is the viscosity solution of (HJB) with $\gamma = 0$. This approach to viscosity solutions is called the method of vanishing viscosity [CL83, CS].

References for Section 8.

[CS] P. Cannarsa, C. Sinestrari, Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004.

[CL83] Crandall M. G., Lions, P. L. (1983) Viscosity solutions of Hamilton-Jacobi equations. Transactions of the American mathematical society, 277(1), 1-42.

[E] L.C. Evans, Lecture notes of the course "An Introduction to Mathematical Optimal Control Theory", https://math.berkeley.edu/ evans/control.course.pdf

[E-SDE] L.C. Evans, Lecture notes of the course "An introduction to stochastic differential equations."

[FR] Fleming W.H., Rishel, R. W. (2012). Deterministic and stochastic optimal control (Vol. 1). Springer Science & Business Media.

[GS-C] Gihman, I. I., Skorohod, A. V. (2012). Controlled stochastic processes. Springer Science & Business Media.

[O] Øksendal, B. Stochastic differential equations. Springer, 2003.

9 Existence and uniqueness of viscosity solutions (continuation of Section 6.1).

Let us recall the statement of the existence and uniqueness theorem for viscosity solutions.

We assume (H0), (H1), $f \in C_{\rm loc}(\mathbb{R}^n \times \mathbb{R}^m)$, $\varphi \in C_{\rm loc}(\mathbb{R}^n)$, and $J_{t,y}(u) = J(u) = \varphi(x(T))$. So we consider the Mayer problem.

Recall that Hamiltonian function for $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ for Mayer OCP is defined by

$$\mathcal{H}(x,p) := \max_{u\in U}(-p \cdot f(x,u)) = \max_{u\in U}\langle -p, f(x,u)\rangle_{\mathbb{R}^n}.$$

The Hamilton-Jacobi-Bellman equation associated with MOCP is

$$-\partial_t v(t,y) + \mathcal{H}(y, \partial_y v(t,y)) = 0, \qquad t \in (0,T), \tag{HJB}$$

where $\partial_t v = \partial v/\partial t$, and $\partial_y v = (\partial v/\partial y_1; \partial v/\partial y_2; \dots, \partial v/\partial y_n)$. To get a boundary value problem for (HJB), we equip it with the terminal value condition

$$v(T,y) = \varphi(y). \tag{TVC}$$

Recall that the value function $V(t,y) = \inf_{u\in L^\infty_U} J_{t,y}(u)$ satisfies (TVC) by definition.

Theorem. *Assume that $\varphi \in \mathrm{Lip}_{\rm loc}(\mathbb{R}^n)$. Then the value function V is a unique viscosity solution to the problem (HJB), (TVC).*

The first condition for a function v to be a viscosity solution is that it should be continuous.

Lemma (regularity)**.** *Let $\varphi \in \mathrm{Lip}_{\rm loc}(\mathbb{R}^n)$. Then $V \in \mathrm{Lip}_{\rm loc}([0,T]\times\mathbb{R}^n)$.*

Proof. We have to prove that for any $r > 0$, there exists C_r s.t.

$$|V(t,y) - V(\widetilde{t},\widetilde{y})| \le C_r(|t-\widetilde{t}| + |y-\widetilde{y}|) \quad \forall t \in [0,T],\ y,\widetilde{y} \in B_r. \tag{LipV}$$

Let us fix $r > 0$. By Lemma UBx from Section 4.2, there exists $R_0 > r$ s.t.

$$|x^{t,y,u}(s)| < R_0 \quad \forall y \in B_r,\ 0 \le t \le s \le T,\ u(\cdot) \in L^\infty_U[0,T].$$

There exists $R > R_0$ s.t.

$$|x^{t,y,u}(s)| < R \quad \forall y \in B_{R_0},\ 0 \le t \le s \le T,\ u(\cdot) \in L^\infty_U[0,T].$$

Since $\varphi \in \mathrm{Lip}_{\rm loc}(\mathbb{R}^n)$, there exists K_φ s.t.

$$|\varphi(x) - \varphi(\widetilde{x})| \le K_\varphi |x - \widetilde{x}| \quad \forall x,\widetilde{x} \in B_R. \tag{Lipφ}$$

Let $M_f := \sup\{|f(x,u)| : x \in B_R, u \in U\}$. Then $M_f < \infty$. By Lemma on uniform continuity from Section 4.2,

$$|x^{t,y,u}(T) - x^{t,\widetilde{y},u}(T)| \le c|y-\widetilde{y}| \qquad \forall t \in [0,T],\ y,\widetilde{y} \in \mathbb{R}^n, u \in L^\infty_U \tag{UCx}$$

with a certain constant $c \ge 1$.

Step 1. Let us consider the case $t = \widetilde{t} \in [0, T)$, take $y, \widetilde{y} \in B_{R_0}$, and prove (LipV). For $t = T$, (LipV) is obvious from (Lipφ) and $V(y, T) = \varphi(T)$.

Assume $V(t, y) \le V(t, \widetilde{y})$. For any $\varepsilon > 0$, there exists a control $u(\cdot)$ s.t.

$$\varphi(x^{t,y,u}(T)) \le V(t, y) + \varepsilon.$$

By the definition of V, (Lipφ), and (UCx),

$$V(t, \widetilde{y}) \le \varphi(x^{t,\widetilde{y},u}(T)) \le \varphi(x^{t,\widetilde{y},u}(T)) + K_\varphi |x^{t,y,u}(T) - x^{t,\widetilde{y},u}(T)| \le V(t, y) + \varepsilon + cK_\varphi |y - \widetilde{y}|.$$

Since ε is arbitrary, we get

$$|V(t, \widetilde{y}) - V(t, y)| \le cK_\varphi |y - \widetilde{y}| \quad \forall y, \widetilde{y} \in B_{R_0},\ t \in [0, T]. \tag{Ly}$$

Step 2. Let us consider arbitrary $t, \widetilde{t} \in [0, T]$ and $y, \widetilde{y} \in B_r$. One can assume $t < \widetilde{t}$. By (DPP1), for any $\varepsilon > 0$, there exists $u \in L^\infty_U$ s.t. for $y_* = x^{t,y,u}(\widetilde{t})$, we have

$$0 \le V(\widetilde{t}, y_*) - V(t, y) \le \varepsilon.$$

Since $r < R_0$, we can apply (Ly) and get

$$|V(\widetilde{t}, y_*) - V(\widetilde{t}, y)| \le cK_\varphi |y_* - \widetilde{y}| \le cK_\varphi (|y_* - y| + |y - \widetilde{y}|) \le cK_\varphi (M_f |\widetilde{t} - t| + |y - \widetilde{y}|)$$

So

$$|V(\widetilde{t}, \widetilde{y}) - V(t, y)| \le cK_\varphi (M_f |\widetilde{t} - t| + |y - \widetilde{y}|) + \varepsilon.$$

Letting $\varepsilon \to 0$, we get

$$|V(\widetilde{t}, \widetilde{y}) - V(t, y)| \le cK_\varphi M_f (|\widetilde{t} - t| + |y - \widetilde{y}|).$$

□

The uniqueness of solution for (HJB), (TVC) follows from the comparison principle (see [CS]). We leave it without a proof.

Lemma. *Assume that $\varphi \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^n)$. Then the value function V is a viscosity solution to the problem (HJB), (TVC).*

9.1 Proof that V is a viscousity solution.

Let us recall the definition of viscosity solution. Let $A \subset \mathbb{R}^n$ be open. Let $F(\cdot) \in C_{\mathrm{loc}}(A \times \mathbb{R} \times \mathbb{R}^n)$. Consider the equation

$$F(x, v, Dv) = 0, \qquad x \in A \subset \mathbb{R}^n. \tag{GenEq}$$

Definition. A function $v \in C_{\mathrm{loc}}(A)$ is called a viscosity subsolution to (GenEq) if

$$F(x, v(x), p) \le 0 \qquad \text{for all } p \in D^+ v(x) \text{ and for all } x \in A. \tag{SubS}$$

A function $v \in C_{\mathrm{loc}}(A)$ is called a viscosity supersolution to (GenEq) if

$$F(x, v(x), p) \ge 0 \qquad \text{for all } p \in D^- v(x) \text{ and for all } x \in A. \tag{SupS}$$

If v satisfies both (SubS) and (SupS), it is called a viscosity solution in A.

We have continuity of V by the first lemma. Let us check (SubS) and (SupS).

Let $(t^0, y^0) \in (0,T) \times \mathbb{R}^n$ and take $R_0 > |y^0|$.

Step 1. Let us prove that V is a viscosity subsolution.

Let $w \in U$ be a constant control and $x(\cdot) = x^{t,y^0,w}(\cdot)$. Then, using (H1), we get

$$x(s) = y^0 + \int_{t^0}^{t} f(x(\tau), w)\mathrm{d}\tau = y^0 + f(y,w)(t-t^0) + \int_{t^0}^{t} [f(x(\tau),w) - f(y,w)]\mathrm{d}\tau = \\ = y^0 + f(y,w)(t-t^0) + o(t-t^0)$$

as $t \searrow t^0$. Let $p = (p^t, p^y) = (p^t, p^y_1, \ldots, p^y_n) \in D^+V(t,y)$.

Recall that $D^+g(z^0)$ is the subdifferential of g at z^0. For $g \in C_{\mathrm{loc}}$, $p \in D^+g(z^0)$ if and only if $p = D\psi(z^0)$ for some $\psi \in C^1_{\mathrm{loc}}$ touching g from above at z^0.

Put $z = (t,y)$ and $z^0 = (t^0, y^0)$. Then near z^0 we have

$$V(z) \le \psi(z) = V(z^0) + \langle p, z - z^0\rangle_{\mathbb{R}^{1+n}} + o(z - z^0).$$

Taking $z = (t, x(t))$ as $t \searrow t^0$, we get

$$V(t,x(t)) \le V(t^0,y^0) + (p^t + f(y^0,w)\cdot p^y)(t-t^0) + o(t-t^0).$$

By the corollary from DPP, V is nondecreasing along the trajectory $(t,x(t))$, $t > t_0$. Hence, $V(t,x(t)) \ge V(t^0,y^0)$ and

$$-p^t - f(y^0,w)\cdot p^y \le 0.$$

Since $w \in U$ is arbitrary,

$$-p^t + \max_{w\in U}(-f(y^0,w)\cdot p^y) = -p^t + \mathcal{H}(y^0,p^y) \le 0.$$

Thus, V is viscosity subsolution to (HJB).

Step 2. Let us prove that V is a viscosity supersolution.

Let B_R and M_f be as before. Then for any control $u(\cdot)$ and $x(t) = x^{t,y^0,u}(t)$,

$$|x(t) - y^0| \le M_f(t-t^0), \quad \forall t > t^0. \tag{*}$$

Let $p = (p^t, p^y) \in D^-V(t,y)$ and $\varepsilon > 0$. From the definition of superdifferential,

$$D^-g(z^0) = \left\{ p \in \mathbb{R}^{1+n} \ : \ \liminf_{z\to z^0} \frac{g(z) - g(z^0) - \langle p, z - z^0\rangle}{|z - z^0|} \ge 0 \right\},$$

we get

$$\frac{V(t,x(t)) - V(t^0,y^0) - p^t(t-t^0) - p^y\cdot(x(t)-y^0)}{|(t,x(t)) - z^0|} \ge -\varepsilon$$

for small enough $|t - t_0|$. Since $|(t,x(t)) - z^0| \ge t - t_0 > 0$, we get

$$\frac{V(t,x(t)) - V(t^0,y^0) - p^t(t-t^0) - p^y\cdot(x(t)-y^0)}{t - t_0} \ge -\varepsilon. \tag{**}$$

By (*) and (H1),

$$p^y \cdot (x(t) - y^0) = \int_{t^0}^{t} p^y \cdot f(x(\tau), u(\tau)) \mathrm{d}\tau \geq$$
$$\geq \int_{t^0}^{t} p^y \cdot f(y^0, u(\tau)) \mathrm{d}\tau - (t - t^0)^2 |p^y| M_f K_1,$$

where K_1 is the Lipschitz constant for f. Hence,

$$p^y \cdot (x(t) - y^0) \geq -(t - t^0)\mathcal{H}(y^0, p^y) - (t - t^0)^2 |p^y| M_f K_1.$$

Combining with (**),

$$\frac{V(t, x(t)) - V(t, x)}{t - t_0} \geq p^t - \mathcal{H}(y^0, p^y) - (t - t^0)|p^y| M_f K_1 - \varepsilon.$$

For any small enough $\delta = t - t^0$, (DPP1) implies that we can choose $u(\cdot)$ and the corresponding controlled trajectory $x(\cdot)$ s.t. $V(t, x(t)) - V(t^0, y^0) \leq \delta^2$. Therefore,

$$p^t - \mathcal{H}(y^0, p^y) \leq \delta(1 + |p^y| M_f K_1) + \varepsilon.$$

Letting $\varepsilon, \delta \to 0$, we get $-p^t + \mathcal{H}(y^0, p^y) \geq 0$. Thus, V is a viscosity supersolution.

From the definition, we see that V is a viscosity solution.

References for Section 9.

[CS] Cannarsa, P., Sinestrari, C., Semiconcave functions, Hamilton-Jacobi equations, and optimal control. Springer, 2004 (Section 7.2).

10 Application to Spectral Optimization. Resonances and their distribution.

Resonances (in the narrow sense) are the frequencies k that describe eigenoscillations of open systems governed by a certain wave equation. These frequencies are, generally, complex numbers $k \in \mathbb{C}$. In many cases, $k \in \mathbb{C}_- := \{z \in \mathbb{C} : \operatorname{Im} z < 0\}$ and either $(-\operatorname{Im} k)$, or $(-\operatorname{Im} k^2)$ characterizes in a certain way the rate of decay of eigenoscillations. The decay is naturally explained by the leakage of the energy of the system to the outer environment.

10.1 Resonances for Schrödinger operator and for the wave equation with a potential.

We will assume in this subsection that $V(\cdot) \in L^\infty_{\mathbb{R}}(\mathbb{R}^3)$. Consider the Schrödinger operator

$$H_V := -\Delta + V(\cdot),$$

with the potential $V(\cdot)$. The operator is defined on $\operatorname{dom} H_V = W^{2,2}_{\mathbb{C}}(\mathbb{R}^3) = \operatorname{dom} \Delta$ and maps $u \in \operatorname{dom} H_V \subset L^2_{\mathbb{C}}(\mathbb{R}^3)$ to $H_V u \in L^2_{\mathbb{C}}(\mathbb{R}^3)$, where

$$(H_V u)(x) = -\Delta u(x) + V(x)u(x), \quad x \in \mathbb{R}^3,$$

in the distributional sense. So H_V is an unbounded operator in the Hilbert space $L^2_{\mathbb{C}}(\mathbb{R}^3)$ defined on the domain $\operatorname{dom} H_V \subset L^2_{\mathbb{C}}(\mathbb{R}^3)$, which is narrower than $L^2_{\mathbb{C}}(\mathbb{R}^3)$, however is dense in $L^2_{\mathbb{C}}(\mathbb{R}^3)$.

This operator is self-adjoint $H_V = H_V^*$ (for basic facts of Operator Theory, see e.g. [RSII, RSIV]). For $\lambda \notin \mathbb{R}$ there exists the inverse operator $(H_V - \lambda)^{-1}$ and this operator is defined on whole $L^2_{\mathbb{C}}(\mathbb{R}^3)$ and bounded. In particular,

$$(H_V - \lambda)(H_V - \lambda)^{-1} f = f \quad \forall f \in L^2_{\mathbb{C}}(\mathbb{R}^3).$$

The operator valued function $(H_V - \lambda)^{-1}$ depending on the complex number λ is called the resolvent of H_V.

Since in the sequel we associate H_V with the wave equation (having a potential term)

$$\partial_t^2 w(t,x) + H_V w(t,x) = 0, \tag{WEq}$$

we write $\lambda = z^2$ and consider the analytic operator-valued function

$$R_V(z) := (H_V - z^2)^{-1}, \qquad k \in \mathbb{C}_+ \setminus \mathrm{i}\mathbb{R}_+,$$

which is the resolvent of H_V taken for the spectral parameter $\lambda = k^2$. Here

$$\mathbb{C}_+ = \{\operatorname{Im} z > 0\}, \quad \mathrm{i}\mathbb{R}_+ = \{\mathrm{i}c : c \in (0, +\infty)\}.$$

Assume that

$$V(x) = 0 \text{ for } x \notin B_{r_0} \text{ with a certain } r_0 > 0. \tag{*}$$

Let $\chi(\cdot) = \chi_{B_{r_1}}(\cdot)$ be the characteristic function of a certain ball B_{r_1} with $r_1 > r_0$, i.e.,

$$\chi(x) = \begin{cases} 1 & \text{if } |x| < r_1 \\ 0, & \text{if } |x| \geq r_1 \end{cases}$$

Theorem (see e.g. [Z17]). *The cut-off resolvent $\chi R_V(z)\chi$, which, for each $z \in \mathbb{C}_+ \setminus \mathrm{i}\mathbb{R}$, is a bounded operator on $L^2_{\mathbb{C}}(\mathbb{R}^3)$ defined by*

$$(\chi R_V(z)\chi u)(x) := \chi(x)R_V(z)\chi(x)u(x),$$

can be continued from $\mathbb{C}_+ \setminus \mathrm{i}\mathbb{R}_+$ to a meromorphic in $\mathbb{C}$ operator-valued function $R_{cont}(z)$, which has at most countable number of poles in $\mathrm{i}\mathbb{R} \cup \overline{\mathbb{C}_-}$.

Definition. The set $\Sigma(H_V)$ of the poles k of this continuation is called the set of resonances.

Remark.

1) The set $\Sigma(H_V)$ of resonances of H_V does not depend on the choice of $r_1 > r_0$ [SZ91].

2) $\Sigma(H_V) \subset \overline{\mathbb{C}_-} \cup \mathrm{i}\mathbb{R}$, and every point of $\Sigma(H_V)$ is isolated (∞ is the only possible accumulation point of $\Sigma(H_V)$).

3) Similarly to eigenvalues, resonances of H_V have finite geometric and algebraic multiplicities. The general definitions of those are somewhat complicated, see remarks in [Z17] and the remarks and references in [K14].

The 'physical meaning' of resonances can be seen from the resonance expansions of scattered waves, for which we follow the recent review paper [Z17]. Let us introduce simplifying technical assumptions

$$\Sigma(H_V) \cap \mathrm{i}[0, +\infty) = \varnothing; \tag{A1}$$

$$\text{all resonances are simple, i.e., have algebraic multiplicity 1;} \tag{A2}$$

$$f \text{ and } g \text{ are certain } L^2_{\mathbb{C}}(\mathbb{R}^3)\text{-functions such that } f(x) = g(x) = 0 \text{ for a.a. } x \notin B_{r_0}. \tag{A3}$$

Theorem (resonance expansion). *Assume (A1)-(A3) and additionally that $f \in W^{1,2}_{\mathbb{C}}(\mathbb{R}^3)$. Let $w(t,x)$, $(t,x) \in \mathbb{R}_+ \times \mathbb{R}^3$, be the solution (in any reasonable sense, classical, weak, or strong) to the wave equation (WEq) satisfying the initial conditions*

$$w(0,\cdot) = f(\cdot), \quad \partial_t w(0,\cdot) = g(\cdot).$$

Then for any $a > 0$, the set $\{k \in \Sigma(H_V) : \operatorname{Im} k > -a\}$ is finite and the following expansion holds

$$w(t,x) = \sum_{\substack{k \in \Sigma(H_V) \\ \operatorname{Im} k > -a}} e^{-\mathrm{i}kt} w_j(x) + b_a(t,x),$$

with certain functions $w_j(\cdot) \in L^2_{\mathbb{C},\mathrm{loc}}(\mathbb{R}^3)$ and the remainder term b_a satisfying

$$\|b_a(t,\cdot)\|_{W^{1,2}(B_r)} \le C_{r,a} e^{-at}(\|f\|_{W^{1,2}(\mathbb{R}^3)} + \|g\|_{L^2(\mathbb{R}^3)}), \quad t > 0,$$

for any $r > 0$ with a certain constant $C_{r,a}$ depending on r and a.

Lemma (elementary properties of $\Sigma(H_V)$).

1) $\Sigma(H_V)$ is symmetric w.r.t. $\mathrm{i}\mathbb{R}$.

2) If $V \equiv 0$, $\Sigma(H_V) = \varnothing$.

3) If $k \in \mathrm{i}\mathbb{R}_+$ is a resonance, then $k^2 \in \mathbb{R}_-$ is a negative eigenvalue of H_V.

4) If $V \geq 0$, then $\Sigma(H_V) \subset \overline{\mathbb{C}_-}$.

Proof of (1) and (2). (1) follows from the properties of the resolvent. Indeed, the adjoint of $(H_V - \lambda)^{-1}$ is

$$((H_V - \lambda)^{-1})^* = ((H_V - \lambda)^*)^{-1} = (H_V^* - \overline{\lambda})^{-1} = (H_V - \overline{\lambda})^{-1},$$

where $\overline{\lambda}$ is the complex conjugate of $\lambda \in \mathbb{C} \setminus \mathbb{R}$. Hence, for $z \in \mathbb{C}_+ \setminus \mathrm{i}\mathbb{R}$, $R_V(z) = \overline{R_V(\widetilde{z})}$, where $\widetilde{z} = -\overline{z}$. This property can be obviously carried over to the cut-off resolvent $\chi R_V(z)\chi$ first for all $z \in \mathbb{C}_+ \setminus \mathrm{i}\mathbb{R}$, and then to its meromorphic continuation $R_{cont}(\cdot)$ to $\mathbb{C}$. So if k is a pole of $R_{cont}(\cdot)$, then $\widetilde{k} = -\overline{k}$ is also a pole of $R_{cont}(\cdot)$. This gives (1).

(2) The explicit form of $R_0(z) = (-\Delta - z^2)^{-1}$ for $z \in \mathbb{C}_+$ is known from the theory of PDE. Namely,

$$(R_0(z)f)(x) = \int G_z(x - x')f(x')\mathrm{d}x', \qquad x \in \mathbb{R}^3,$$

where

$$G_z(x) := \frac{e^{\mathrm{i}z|x|}}{4\pi|x|}, \qquad x \neq 0.$$

Note that $G_z(x)$ is analytic in $\mathbb{C}$ function in the variable z for every $x \neq 0$. Applying the cut-off procedure and continuing $R_{cont}(\cdot)$ to $\mathbb{C}$, one obtains (2). □

The proof of (3) requires some basic knowledge of operator theory [RSII, RSIV]. The resolvent $(H-\lambda)^{-1}$ of an operator H in a Hilbert space is the operator-valued function defined on the set $\rho(H)$ that consists of $\lambda \in \mathbb{C}$ such that that $H - \lambda$ is invertible and $(H - \lambda)^{-1}$ is a bounded operator defined on the whole Hilbert space. So the function $(H - \cdot)^{-1}$ acts from $\rho(H)$ to the Banach space of bounded linear operators. This operator-valued function is analytic on $\rho(H)$. The spectrum $\sigma(H)$ is by the definition $\sigma(H) = \mathbb{C} \setminus \rho(H)$. If λ is an eigenvalue of H, then $\lambda \in \sigma(H)$ (generally, a point of spectrum is not necessarily an eigenvalue of H). If $H = H^*$, then $\sigma(H) \subset \mathbb{R}$. If $V(x) \geq c \in \mathbb{R}$ a.e., then $\sigma(H_V) \subset [c, +\infty)$.

Assume now that (*) is valid. Then $\sigma(H_V) \supset [0, +\infty)$ and $\sigma(H_V) \cap \mathbb{R}_-$ consists of a finite number $n \in \mathbb{N} \cup \{0\}$ of isolated eigenvalues $\{\lambda_j\}_{j=1}^n$ (if $V \geq 0$, then $n = 0$). Moreover, each eigenvalue has finite algebraic multiplicity.

Remark. For selfadjoint operators, geometric and algebraic multiplicities of eigenvalues coincide. For resonances, generally, this is not true. Roughly speaking, resonances are eigenvalues of certain nonselfadjoint operators associated with H_V in a special way. For a resonance $k \in \mathbb{C}_-$, nontrivial root subspaces can appear, and then algebraic multiplicity is greater than geometric multiplicity.

If $z_0 \in \mathbb{C}_+$ is such that $z_0^2 \notin \{\lambda_j\}_{j=1}^n$, then $R_V(z)$ and $R_{cont}(z)$ are analytic at z_0 and so $z_0 \notin \Sigma(H_V)$. The spectral decomposition of $(H_V - \lambda)^{-1}$ near an isolated eigenvalue λ_j implies that λ_j is a pole of $(H_V - \lambda)^{-1}$ and that the corresponding $k \in \mathrm{i}\mathbb{R}_+$ is also a pole for the cut-off resolvent. This implies the statement (3) of the lemma.

The statement (4) of the lemma follows from (3) and (2).

Definition. The set $\sigma_{disc}(H)$ of isolated eigenvalues of an operator H is called the discrete spectrum. The essential spectrum is defined by $\sigma_{ess}(H) := \sigma(H) \setminus \sigma_{disc}(H)$.

Under the assumptions $V \in L^\infty_R$ and (*),

$$\sigma_{disc}(H_V) = \{\lambda_j\}_{j=1}^n, \qquad \sigma_{ess}(H_V) = [0, +\infty).$$

The operator $H_0 = (-\Delta)$ has no eigenvalues and $\sigma(-\Delta) = \sigma_{ess}(-\Delta) = [0, +\infty)$.

Roughly speaking, the resonances are the poles of the resolvent continued through the essential spectrum in the generalized way described above.

10.1.1 More delicate properties of $\Sigma(H_V)$.

Theorem (Rellich's uniqueness theorem)**.** *If $k \in \mathbb{R}$ is a resonance, then $k = 0$.*

Remark. This result is not valid for some wider classes of potentials.

Theorem ([SZ16])**.** *If $V \not\equiv 0$, then $\Sigma(H_V) \neq \varnothing$.*

Open Problem (see Section 2.7 in [Z17]). Prove that $V \not\equiv 0$ implies $\#\Sigma(H_V) = \infty$.

Here $\#(\Sigma(H_V) \cap E)$ is a number of resonances in the set $E \subset \mathbb{C}$ (taking their algebraic multiplicities into account).

The set $\Sigma(H_V)$ is studied mainly in terms of the counting function

$$\mathcal{N}_{H_V}(r) := \#\{k \in \Sigma(H_V) : |k| \leq r\}.$$

It is known that $\mathcal{N}_{H_V}(r) \leq Cr^3$ for certain constant C depending on $V(\cdot)$ (Zworski, 1991).

Conjecture (see Section 2.7 in [Z17] and the open problem above). Assume that $V \not\equiv 0$ (in L^∞-sense). Then $\mathcal{N}_{H_V}(r) \geq cr^3$ for a certain constant $c > 0$.

Except radially symmetric cases, the presently available information about the structure of $\Sigma(H_V)$ is very limited.

10.2 Resonances of point interactions and examples of asymptotic sequences.

Let us consider, following [AGHH], Schrödinger operators $H_{a,Y}$ with point interaction, which are associated with the formal differential expression

$$-\Delta u(x) + \text{“}\sum_{j=1}^N \mu(a_j)\delta(x - y_j)u(x)\text{“}, \quad x \in \mathbb{R}^3,\ N \in \mathbb{N}, \tag{PI}$$

where $\delta(\cdot)$ is the Dirac δ-function. The distinct interaction centers $y_j \in \mathbb{R}^3$ form the finite family $Y = \{y_j\}_{j=1}^N$, and $a = (a_j)_{j=1}^N \in \mathbb{C}^N$ is the tuple of 'strength' parameters.

The simplest way to define rigorously the operator $H_{a,Y}$ is via its resolvent. Consider the operator-valued function $R_{a,Y}(z)$,

$$(R_{a,Y}(z)f)(\cdot) = \int K_{a,Y}(\cdot, x')f(x')\mathrm{d}x', \qquad f \in L^2_{\mathbb{C}}(\mathbb{R}^3),$$

defined for $z \in \mathbb{C}_+ \setminus \mathrm{i}\mathbb{R}_+$ by its integral kernel

$$K_{a,Y}(x,x') = G_z(x-x') + \sum_{j,j'=1}^{N} G_z(x-y_j)\,[\Gamma_{a,Y}]^{-1}_{j,j'}\,G_z(x'-y_{j'}), \qquad (**)$$

where

- $x, x' \in \mathbb{R}^3 \setminus Y$, $x \neq x'$,
- $G_z(x-x') := \frac{e^{\mathrm{i}z|x-x'|}}{4\pi|x-x'|}$ is the integral kernel associated with $(-\Delta - z^2)^{-1}$,
- $[\Gamma_{a,Y}]^{-1}_{j,j'}$ denotes the j,j'-element of the inverse to the matrix

$$\Gamma_{a,Y}(z) = \Big[\big(a_j - \tfrac{\mathrm{i}z}{4\pi}\big)\,\delta_{jj'} - \widetilde{G}_z(y_j - y_{j'})\Big]_{j,j'=1}^{N}, \text{ where } \widetilde{G}_z(x) := \begin{cases} G_z(x), & x \neq 0 \\ 0, & x = 0 \end{cases}.$$

(The notation $\delta_{jj'}$ as usually stands for the Kronecker delta.)

Example. Let Y consist of $N = 2$ point interactions at $y_1, y_2 \in \mathbb{R}^3$. Then

$$\Gamma_{a,Y}(z) = \begin{pmatrix} a_1 - \frac{\mathrm{i}z}{4\pi} & -G_z(y_1-y_2) \\ -G_z(y_1-y_2) & a_2 - \frac{\mathrm{i}z}{4\pi} \end{pmatrix}.$$

One can write the elements $[\Gamma_{a,Y}]^{-1}_{j,j'}$ of the inverse matrix $\Gamma^{-1}_{a,Y}$ explicitly as exponential polynomials.

Theorem. *Let $a = (a_j)_{j=1}^N \in \mathbb{R}^n$. Then there exists a selfajoint in $L^2_{\mathbb{C}}(\mathbb{R}^3)$ operator $H_{a,Y}$ such that $(H_{a,Y} - z^2)^{-1} = R_{a,Y}(z)$ for all $z \in \mathbb{C}_+ \setminus \mathrm{i}\mathbb{R}_+$.*

So the Krein-type formula (**) for the difference of the perturbed and unperturbed resolvents of operators $H_{a,Y}$ and $-\Delta$ can be used to define the operator $H_{a,Y}$ with point interactions.

Definition 2 [AH84]. The set of resonances $\Sigma(H_{a,Y})$ associated with the operator $H_{a,Y}$ (in short, resonances of $H_{a,Y}$) is by definition the set of $k \in \mathbb{C}$ such that $\det \Gamma_{a,Y}(k) = 0$.

One can consider also Definition 1 for this class of the operators. It is easy to see that these two definitions are equivalent.

Remark. The multiplicity of k is by the definition [AH84] its multiplicity as the zero of $\det \Gamma_{a,Y}(\cdot)$. This multiplicity should coincide with algebraic multiplicity of k in the sense of [Z17], but I am not sure if this have been carefully checked by someone. Some care is usually needed with the multiplicity of 0.

Example ([AH84], see also [AGHH]). Consider the previous example with $N = 2$, $\ell = |y_1 - y_2| > 0$, and $a_1 = a_2 = \alpha \in \mathbb{R}$. Then $\det \Gamma_{a,Y}(k) = 0$ takes the form

$$(4\pi\ell\alpha - \mathrm{i}\ell k)^2 - e^{\mathrm{i}2\ell k} = 0.$$

The solutions to this equation can be written with the use of zeroes of elementary (but transcendental) functions. In particular, it can be shown that

$$\Sigma_{H_{a,Y}} = \{k_n^+\}_{n\in\mathbb{N}} \cup \{k_n^-\}_{n\in\mathbb{N}},$$

where

$$k_n^+ = (n-1/2)\frac{\pi}{\ell} - \frac{\mathrm{i}}{\ell}\ln((n-1/2)\pi) + o(1) \text{ as } n \to +\infty.$$

To obtain asymptotics of k_n^-, it is enough to replace n to $(-n)$ in the above formula.

10.3 Asymptotic structure of the set of resonances.*

Example ([AK18]). Consider the case $N = 2$, $|y_j - y'_j| = 1$ for all $j \neq j'$, and $a_1 = a_2 = a_3 = 0$. Then it can be shown that

$$\Sigma_{H_{a,Y}} = \bigcup_{j=1}^{3} \{k_{j,n}^{+}\}_{n=n_j^{+}}^{+\infty} \cup \bigcup_{j=1}^{3} \{k_j^{-}\}_{n=n_j^{-}}^{+\infty}$$

with certain $n_j^{\pm} \in \mathbb{Z}$, where

$$k_{j,n}^{+} = 2\pi n - \mathrm{i}\ln n + \pi/2 - \mathrm{i}\ln(2\pi) + o(1) \text{ as } n \to +\infty \quad \text{ for } j = 1, 2$$

and

$$k_{3,n}^{+} = 2\pi n - \mathrm{i}\ln n - \pi/2 - \mathrm{i}\ln(\pi) + o(1) \text{ as } n \to +\infty,$$

The asymptotics of $k_{j,n}^{-}$ can be obtained by the reflection w.r.t. $\mathrm{i}\mathbb{R}$.

Remark. It is shown in [AK18] that for arbitrary $H_{a,Y}$ with $N \geq 2$, $\Sigma(H_{a,Y})$ can be decomposed into $N_1 \leq N$ of sequences with asymptotics

$$2\pi\mu_m n - \mathrm{i}\mu_m \ln|n| + O(1) \text{ as } |n| \to +\infty, \qquad m = 1, \ldots, N_1,$$

with certain $\mu_m > 0$. These leading parameters μ_m are connected with the mutual placement of the centers y_j, but in quite a nontrivial way.

References for Section 10.

[AGHH] S. Albeverio, F. Gesztesy, R. Høegh-Krohn, H. Holden, Solvable models in quantum mechanics. 2nd edition, with an appendix by P. Exner. AMS Chelsea Publishing, Providence, RI, 2012.

[AH84] S. Albeverio, R. Høegh-Krohn, Perturbation of resonances in quantum mechanics. J. Math. Anal. Appl. 101 (1984), 491–513.

[AK18] Albeverio, S. and Karabash, I.M., 2018. On the multilevel internal structure of the asymptotic distribution of resonances. arXiv preprint arXiv:1807.02889.

[K14] I.M. Karabash, Pareto optimal structures producing resonances of minimal decay under L^1-type constraints, *Journal of Differential Equations* **257** (2014), no.2, 374–414.

[RSII] Reed, M., Simon, B. (1975). Methods of modern mathematical physics II: Fourier Analysis, Self-Adjointness (Vol. 2).

[RSIV] Reed, M., Simon, B. (1978). Methods of modern mathematical physics IV: Analysis of Operators.

[SZ91] Sjostrand, J. and Zworski, M., 1991. Complex scaling and the distribution of scattering poles. Journal of the American Mathematical Society, 4(4), pp.729-769.

[SZ16] Smith, H., Zworski, M.: Heat traces and existence of scattering resonances for bounded potentials. Ann. Inst. Fourier 66, 455–475 (2016)

[Z17] Zworski, M. (2017). Mathematical study of scattering resonances. Bulletin of Mathematical Sciences, 7(1), 1-85.

11 Application to Spectral Optimization. Pareto optimization of resonances.*

11.1 Resonances in layered optical cavity

For the idealized model involving a layered optical cavity and normally passing electromagnetic waves, the Maxwell system can be reduced to the wave equation of a nonhomogeneous string

$$\varepsilon(s)\partial_t^2 v(s,t) = \partial_s^2 v(s,t), \qquad s \in \mathbb{R}, \tag{WEq}$$

where v is one of the normal components of the electric field and $\varepsilon(\cdot)$ is the spatially varying dielectric permittivity of layers.

So, for a multilayer cavity, $\varepsilon(\cdot)$ is a step function that can take several values $\widehat{\epsilon}_1, \dots, \widehat{\epsilon}_m > 0$, corresponding to the materials available for fabrication. In a finite interval $s \in [s_-, s_+]$, the function $\varepsilon(\cdot)$ represents the nonhomogeneous structure of the resonator ('nonhomogeneous' means that, generally, $\varepsilon(\cdot)$ is not necessarily a constant on the whole interval). Outside of this interval $\varepsilon(\cdot)$ equals to the constant permittivity $\epsilon_\infty = \mathfrak{n}_\infty^2$ (where $\mathfrak{n}_\infty > 0$ is the corresponding refractive index) of the homogeneous outer medium

$$\varepsilon(s) \equiv \epsilon_\infty > 0 \text{ for all } s \in \mathbb{R} \setminus [s_-, s_+].$$

Mathematically, it is convenient to consider a wider class of resonators $\varepsilon(\cdot)$ assuming that

$$\varepsilon \in L^\infty(s_-, s_+) \text{ and } \varepsilon(s) > 0 \text{ almost everywhere (a.e.).}$$

Resonances of the cavity can be defined in several equivalent ways. One way is via poles of the meromorphic extension of the cut-off version of the resolvent $\left(-\frac{1}{\varepsilon(s)}\partial_s^2 - z^2\right)^{-1}$, where the extension is done from the upper complex half-plane $\mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ to the whole plane $\mathbb{C}$. In 1-dim. case, there is a simpler definition via a generalized eigenvalue problem with the eigen-parameter entering into the boundary conditions.

Let $s_\pm$ be fixed, $-\infty < s_- < s_+ < +\infty$.

Definition. A resonance associated with the function $\varepsilon(s)$, $s \in (s_-, s_+)$, is a number $k \in \mathbb{C} \setminus \{0\}$ such that the (generalized) eigenproblem

$$y''(s) = -k^2\varepsilon(s)y(s) \quad \text{for a.a. } s \in \mathbb{R}, \tag{Eq}$$
$$y'(s_\pm) = \pm \mathrm{i}k\mathfrak{n}_\infty\, y(s_\pm) \tag{BC$_\pm$}$$

has a nontrivial solution $y \in W_{\mathbb{C}}^{2,\infty}[s_-, s_+]$ (nontrivial means that y is not identically 0 in the $L^\infty(\mathbb{R})$-sense). Such a solution y is called a (resonant) *mode* associated with k and $\varepsilon(\cdot)$. The set of all nonzero resonances of $\varepsilon(\cdot)$ is denoted by $\Sigma(\varepsilon)$.

Remark. This definition is not completely equivalent to earlier definitions because we *explicitly exclude* 0 from the set of resonances $\Sigma(\varepsilon)$. The reason for this is that for the particular case of the problem (Eq)-(BC$_\pm$), for any $\varepsilon(\cdot)$, the value of $k = 0$ corresponds to the nontrivial solution $y \equiv 1$. This solution in the present case is not interesting for applications.

The real part $\operatorname{Re} k$ and the negative $(-\operatorname{Im} k)$ of the imaginary part of k in the context of the wave equation (WEq) corresponds to the (real angular) *frequency* of eigenoscillations $e^{-ikt}y(s)$, the negative $(-\operatorname{Im} k)$ of the imaginary part to the (exponential) *decay rate*. Therefore, for the pairs $(k;\varepsilon(\cdot))$ s.t. $k \in \Sigma(\varepsilon)$, we introduce the decay rate functional

$$\operatorname{Dr}(k;\varepsilon) := -\operatorname{Im} k.$$

The value $(-2\operatorname{Im} k)$ is called the *bandwidth* of a resonance k.

Every (nonzero) resonance k has $\operatorname{Im} k < 0$. That is, $\Sigma(\varepsilon) \subset \mathbb{C}_- := \{\operatorname{Im} z < 0\}$ and $\operatorname{Dr}(k,\varepsilon) > 0$.

11.2 Simplified statements of the problem of optimization of resonances

Let

$$\mathbb{F}_{s_-,s_+} := \{\varepsilon(\cdot) \in L^\infty_{\mathbb{R}}(s_-,s_+) \ : \ \mathfrak{n}_1^2 \le \varepsilon(s) \le \mathfrak{n}_2^2 \text{ for a.a. } s \in (s_-,s_+)\},$$

where $0 < \mathfrak{n}_1 < \mathfrak{n}_2$.

Recall that under the condition $k \in \Sigma(\varepsilon)$, the decay rate functional is $\operatorname{Dr}(k;\varepsilon) = -\operatorname{Im} k$.

Simplified statement of the problem of minimization of decay rate.

Find

$$\inf_{\substack{k\in\Sigma(H_\varepsilon)\\ \varepsilon\in\mathbb{F}_{s_-,s_+}}} \operatorname{Dr}(k;\varepsilon) \qquad \text{and} \qquad \arg\min_{\substack{k\in\Sigma(H_\varepsilon)\\ \varepsilon\in\mathbb{F}_{s_-,s_+}}} \operatorname{Dr}(k;\varepsilon) \tag{P0}$$

.

Problems of such type were introduced for Schrödinger operators $-\Delta + V(\cdot)$ by [HS86]. For layered optical cavities, the problem

$$\arg\min_{\substack{k\in\Sigma(H_\varepsilon)\\ \varepsilon\in\mathbb{F}_{s_-,s_+}}} Q(k;\varepsilon)$$

with $Q(k,\varepsilon) = \frac{|\operatorname{Re} k|}{-2\operatorname{Im} k}$ was considered numerically by [KS08] with the use of steepest ascent method. The quantity $Q = \frac{|\operatorname{Re} k|}{-2\operatorname{Im} k}$ is called the quality-factor (Q-factor) associated with a resonance k of the cavity $\varepsilon(\cdot)$. A problem very close to (P0) was considered numerically by steepest ascent method in [HBKW08].

In comparison with more classical optimization problems for *eigenvalues* $\lambda \in \mathbb{R}$ of self-adjoint opeartors, two new theoretical and numerical difficulties appear for the two resonance optimization problems mentioned above:

(a) it is difficult to prove existence of optimizers;

(b) resonances may have algebraic multiplicity > 1, and then, they are nondifferentiable w.r.t. $\varepsilon(\cdot)$ [K14] (this creates difficulties for the gradient methods and for the necessary conditions of optimality).

11.3 Pareto optimization of resonances

For general theory of Pareto optimization, see [BV]. Our definitions are slightly different, but the main idea, which goes back to Vilfredo Pareto, is the same. (Vilfredo Pareto (1848–1923) was one of the founders of Mathematical Economics).

This and the next subsections give definitions and results of [K13, K14, KLV17-1].

Problem of minimization of decay rate for particular frequencies. Let $\alpha \in \mathbb{R}$. Find

$$\beta_{\min}(\alpha) := \inf_{\substack{k\in\Sigma(\varepsilon)\\ \operatorname{Re} k=\alpha\\ \varepsilon\in\mathbb{F}_{s_-,s_+}}} (-\operatorname{Im} k) \quad \text{and} \quad \operatorname*{arg\,min}_{\substack{k\in\Sigma(\varepsilon)\\ \operatorname{Re} k=\alpha\\ \varepsilon\in\mathbb{F}_{s_-,s_+}}} \operatorname{Dr}(k;\varepsilon).$$

The value $\beta_{\min}(\alpha)$ is called the minimal decay rate for the frequency α. Let us denote by $\operatorname{dom}\beta_{\min} := \{\alpha \in \mathbb{R} : \beta_{\min}(\alpha) < +\infty\}$ (recall that $\inf \varnothing = +\infty$). The set $\operatorname{dom}\beta_{\min}$ is called the set of achievable frequencies.

Definition. The set

$$P := \{\alpha - \mathrm{i}\beta_{\min}(\alpha) \ : \ \alpha \in \operatorname{dom}\beta_{\min}\},$$

is *Pareto optimal frontier* of minimal decay.

Let us define *set of achievable resonances* by

$$\Sigma[\mathbb{F}_{s_-,s_+}] := \bigcup_{\varepsilon\in\mathbb{F}_{s_-,s_+}} \Sigma(\varepsilon)$$

Then $\Sigma[\mathbb{F}_{s_-,s_+}] \subset \mathbb{C}_-$.

Obviously, P is a subset of the boundary $\partial\Sigma[\mathbb{F}_{s_-,s_+}]$ of $\Sigma[\mathbb{F}_{s_-,s_+}]$.

Reformulation of the problem. We want to find ε that generate resonances k lying on the Pareto optimal frontier.

Other types of definitions of optimizers were introduced in [HS86, K14, KLV17].

Definition. An achievable resonance k_0 is called the *resonance of minimal decay for (the frequency)* $\alpha_0 = \operatorname{Re} k_0$ if k_0 belongs to the *Pareto frontier of minimal decay* P. If $k_0 \in P$, then $\varepsilon \in \mathbb{F}_{s_-,s_+}$ such that $k_0 \in \Sigma(\varepsilon)$ is called the *resonator of minimal decay for* α_0.

This approach easily removes the problem with the existence of optimizers.

Lemma (existence)**.**

1) $\Sigma[\mathbb{F}_{s_-,s_+}]$ *is closed.*

2) For every achievable frequency α*, there exists a resonance of minimal decay* $k = \alpha - \mathrm{i}\beta_{\min}(\alpha)$ *and an associated optimal resonator* $\varepsilon(\cdot) \in \mathbb{F}_{s_-,s_+}$ *generating* k*.*

Proof. (1) can be obtained from weak* compactness of $\mathbb{F}_{s_-,s_+}$ in the way similar to the proof of the existence of an optimizer for Mayer's problem (see Section 4.1 and [HS86, K13, KLV17-1]).

(2) follows from (1) and $P \subset \partial\Sigma[\mathbb{F}_{s_-,s_+}] \subset \Sigma[\mathbb{F}_{s_-,s_+}]$. □

Note that $P \subset \Sigma[\mathbb{F}_{s_-,s_+}] \subset \mathbb{C}_-$, so $\beta_{\min}(\alpha) > 0$ for all $\alpha \in \mathbb{R}$.

Remark. A resonator of minimal decay for a particular frequency α is not necessarily unique. A simple non-uniqueness example for $\alpha = 0$ was constructed in [KLV17-2]. There exists numerical evidence that non-uniqueness may happen also for $\alpha \neq 0$ [KKV18].

Consider the nonlinear equation of the bang-bang type

$$-y''(s) = k^2 y E(y^2(s)), \quad \text{where } E(z) := \begin{cases} \epsilon_2, & z \in \mathbb{C}_+ \\ \epsilon_1, & z \in \mathbb{C}_- \end{cases}. \tag{NEq}$$

Theorem (an analogue of Euler-Lagrange equation, [K13, KLV17]). *Let $\alpha \in \operatorname{dom} \beta_{\min}$ and $k = \alpha - \mathrm{i}\beta_{\min}(\alpha)$ be the resonance of minimal decay for α. Then:*

(1) There exists a nontrivial solution y to the bang-bang boundary value problem (NEq)-($BC_\pm$).

(2) $k \in \Sigma(\varepsilon)$ for a certain $\varepsilon \in \mathbb{F}_{s_-,s_+}$ if and only if there exists a nontrivial solution y to the bang-bang boundary value problem (NEq)-($BC_\pm$) satisfying $\varepsilon(\cdot) = E(y^2(s))$ a.e..

The proof is based on the multi-parameter perturbation theory for resonances [K13, K14].

Remark. 1) It follows from the theorem that each resonator $\varepsilon(\cdot)$ of minimal decay is of bang-bang type in the sense of Control Theory.

2) The requirement that k is a resonance of minimal decay, can be replaced by the condition that $k \in \partial\Sigma[\mathbb{F}_{s_-,s_+}] \setminus \mathrm{i}\mathbb{R}$.

11.4 Symmetric resonators

Most of research for 1-dim. photonic crystals have been done under the assumption that

$$\varepsilon(\cdot) \text{ is symmetric w.r.t. the } \textit{resonator center } s^{\text{centr}} = \tfrac{s_- + s_+}{2}$$

in the sense that $\varepsilon(\cdot - s^{\text{centr}})$ is an even function. For such symmetric $\varepsilon(\cdot)$, the *resonant mode* $y(\cdot)$ is either an even, or odd function w.r.t. s^{centr} (i.e., $y(\cdot - s^{\text{centr}})$ is even, or odd), and therefore satisfies

$$\text{either the condition } y'(s^{\text{centr}}) = 0, \text{ or the condition } y(s^{\text{centr}}) = 0.$$

These conditions can be treated as boundary conditions and can be used to simplify the problem.

Shifting s^{centr} to zero and getting $s_+ = -s_- = d$ with a certain $d > 0$, we introduce the family

$$\mathbb{F}_d^{\text{sym}} = \{\varepsilon \in \mathbb{F}_{-d,d} \ : \ \varepsilon(s) = \varepsilon(-s) \text{ a.e.}\}$$

and, for $\varepsilon \in \mathbb{F}_d^{\text{sym}}$, introduce the set $\Sigma^{\text{even}}(\varepsilon)$ (the set $\Sigma^{\text{odd}}(\varepsilon)$) of resonances k such that the corresponding mode y is an even (resp., odd) function. We will say that the corresponding k is an *even-mode* resonance (resp., *odd-mode* resonance) of $\varepsilon(\cdot)$.

To introduce the optimization problems for odd-mode and even-mode resonances, one can define the sets of achievable even- and odd-mode resonances

$$\Sigma^{\text{even(odd)}}[\mathbb{F}_d^{\text{sym}}] := \bigcup_{\varepsilon \in \mathbb{F}_d^{\text{sym}}} \Sigma^{\text{even(odd)}}(\varepsilon),$$

the corresponding functions $\beta_{\min}^{\text{even(odd)}}(\alpha)$, $\alpha \in \mathbb{R}$, and Pareto optimal frontiers

$$P^{\text{even(odd)}} := \{\alpha - \mathrm{i}\beta_{\min}^{\text{even(odd)}}(\alpha) \ : \ \alpha \in \operatorname{dom} \beta_{\min}^{\text{even(odd)}}\}$$

of even- and odd-mode resonances of minimal decay.

The theorem about 'Euler-Lagrange eigenproblem' is valid for these symmetric optimization problems if an additional condition $y'(0) = 0$ or $y(0) = 0$ is imposed.

References for Section 11.

[BV] S. Boyd, L. Vandenberghe, Convex optimization. Cambridge university press, Cambridge, 2004.

[HS86] E.M. Harrell, R. Svirsky, Potentials producing maximally sharp resonances, *Transactions of the American Mathematical Society* **293(2)** (1986), 723–736.

[HBKW08] P. Heider, D. Berebichez, R.V. Kohn, and M.I. Weinstein, Optimization of scattering resonances, *Struct. Multidisc. Optim.* **36** (2008), 443–456.

[KS08] C.-Y. Kao, F. Santosa, Maximization of the quality factor of an optical resonator, *Wave Motion* **45** (2008), 412–427.

[K13] I.M. Karabash, Optimization of quasi-normal eigenvalues for 1-D wave equations in inhomogeneous media; description of optimal structures, Asymptotic Analysis 81 (2013) no.3-4, 273-295.

[K14] I.M. Karabash, Pareto optimal structures producing resonances of minimal decay under L^1-type constraints, *Journal of Differential Equations* **257** (2014), no.2, 374–414.

[KKV18] I.M. Karabash, H. Koch, I.V. Verbytskyi, Pareto optimization of resonances and minimum-time control, preprint arXiv:1808.09186, https://arxiv.org/pdf/1808.09186

[KLV17-1] Karabash, I. M., Logachova, O. M., Verbytskyi, I. V. Nonlinear Bang–Bang Eigenproblems and Optimization of Resonances in Layered Cavities. Integral Equations and Operator Theory 88(1),(2017), 15-44;

[KLV17-2] Karabash, I. M., Logachova, O. M., Verbytskyi, I. V. Overdamped modes and optimization of resonances in layered cavities, Methods of Functional Analysis and Topology 23 (2017), no.3, 252–260.

12 Application to Spectral Optimization. Minimum-time control and resonators of minimal length.*

12.1 Dual problem of minimization of length of a resonator

We will formulate a dual optimization problem, where the length of resonator becomes a cost function, and the resonance k is fixed. In some sense, the constraints and the cost functions exchange their roles in comparison with the problem (P0) of the previous section. At the end, we will show that under certain constraints the Pareto optimization problem can be reduced to the dual problem of the minimization of the length.

This section follows [KKV18].

To play with the length $s^+ - s^-$ of the resonator, we introduce another feasible family that makes the length a free parameter. Some additional care is needed in the case when $\epsilon_\infty \in [\epsilon_1, \epsilon_2]$ because in this case the definition of the length of resonator is ambiguous.

Recall that $n_j = \epsilon_j^{1/2} > 0$, $j = 1, 2, \infty$, are refractive indices corresponding to the permittivities ϵ_j.

The family $\mathbb{F}$ of feasible (permittivity) coefficients consists, by definition, of positive functions $\varepsilon(\cdot) \in L^\infty_{\mathbb{R}}(\mathbb{R})$ such that there exists $s_\pm \in \mathbb{R}$ satisfying $s_- \le s_+$ and the following conditions:

$$\varepsilon(s) = \mathfrak{n}_\infty^2 \text{ for a.a. } s \in \mathbb{R} \setminus [s_-, s_+],$$
$$\mathfrak{n}_1^2 \le \varepsilon(s) \le \mathfrak{n}_2^2 \text{ for } s \in (s_-, s_+),$$

where $\mathfrak{n}_\infty, \mathfrak{n}_1, \mathfrak{n}_2$ are fixed constants satisfying $0 < \mathfrak{n}_\infty$ and $0 < \mathfrak{n}_1 < \mathfrak{n}_2$.

Definition. For any given $\varepsilon(\cdot)$ that is not equal to the constant function $\mathfrak{n}_\infty^2$, we denote by $[s_-^\varepsilon, s_+^\varepsilon]$ the shortest interval $[s_-, s_+]$ satisfying (12.1), and by $\ell(\varepsilon) := s_+^\varepsilon - s_-^\varepsilon$ *the effective length of the resonator* defined by the coefficient $\varepsilon(\cdot)$. If $\varepsilon(\cdot) = \mathfrak{n}_\infty^2$ (in $L^\infty(\mathbb{R})$-sense), we put $s_-^\varepsilon = s_+^\varepsilon = 0$ and $\ell(\varepsilon) = 0$.

Recall that a resonance of associated with $\varepsilon(\cdot)$ is a number $k \in \mathbb{C}$ such that the (generalized) eigenproblem

$$y''(s) = -k^2 \varepsilon(s) y(s) \quad \text{for } s \in \mathbb{R}, \tag{Eq}$$
$$\frac{y'(s)}{k} = \pm \mathrm{i}\mathfrak{n}_\infty \, y(s) \quad \text{for } s = s_\pm^\varepsilon \tag{BC$_p$m}$$

has a nontrivial solution $y \in W^{2,\infty}_{\mathrm{loc},\mathbb{C}}(\mathbb{R})$ (nontrivial means that y is not identically 0 in the $L^\infty(\mathbb{R})$-sense).

Lemma. *For a nontrivial solution $y(\cdot)$ to (Eq), equality (BC$_\pm$) is satisfied for $s = s_\pm^\varepsilon$ if and only if it is satisfied for certain s such that $\pm s > \pm s_\pm^\varepsilon$.*

Proof. Consider (BC$_+$). Then the both cases are equivalent to the statement that $y(s) = a_+ \exp(\mathrm{i}\mathfrak{n}_\infty \, ks)$ for all $s \ge s_+^\varepsilon$ with a certain constant $a_+ \in \mathbb{C} \setminus \{0\}$. □

Consider the following minimization problem

$$\arg\min_{\substack{\varepsilon \in \mathbb{F} \\ k \in \Sigma(\varepsilon)}} \ell(\varepsilon), \tag{MinL}$$

where the resonance $k \in \mathbb{C}_-$ and the material parameters $\mathfrak{n}_\infty$, $\mathfrak{n}_1$, $\mathfrak{n}_2$ are fixed.

Without a priori knowledge if any *minimizers* $\varepsilon(\cdot)$ for the problem (MinL) exist, one can define the corresponding minimum length

$$\ell_{\min}(k) := \inf_{\substack{\varepsilon \in \mathbb{F} \\ k \in \Sigma(\varepsilon)}} \ell(\varepsilon).$$

12.2 The minimum-time reformulation of the minimization of length.

Let $k \neq 0$ be fixed. Let us interpret s as time and functions $\varepsilon(\cdot) \in L^\infty_{\mathbb{R},\mathrm{loc}}[s_-, +\infty)$ as *control strategies* (or, slightly abusing the terminology, we will say that $\varepsilon(\cdot)$ are controls).

The family $\mathbb{F}_{s_-}$ of *feasible controls* is defined by

$$\mathbb{F}_{s_-} := \{\varepsilon(\cdot) \in L^\infty_{\mathbb{R}}(s_-, +\infty) \ : \ \mathfrak{n}_1^2 \leq \varepsilon(s) \leq \mathfrak{n}_2^2 \text{ for } s > s_-\}.$$

To modify the differential equation (Eq) into a control system in a state-space $\mathbb{C}^2$, we denote $Y_0(s) = y(s)$, $Y_1(s) = \frac{y'(s)}{\mathrm{i}k}$, form a column vector $Y(s) = (Y_0; Y_1)^\top \in \mathbb{C}^2$, and write (Eq) as

$$Y'(s) = \mathrm{i}k \begin{pmatrix} 0 & 1 \\ \varepsilon(s) & 0 \end{pmatrix} Y(s). \tag{YEq}$$

Since this system is linear one can consider the associated dynamics on the complex projective line, which we identify with the Riemann sphere $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. From the point of view of elementary ODEs, this is the standard reduction to the associated Riccati differential equation.

Namely, for a nontrivial solution $y(\cdot)$ to (Eq), the dynamics of the function $x(\cdot)$ defined by

$$x(s) = \frac{y'(s)}{\mathrm{i}k y(s)} \text{ if } y(s) \neq 0, \quad x(s) = \infty \text{ if } y(s) = 0,$$

is described by the control system

$$x'(s) = f(x(s), \varepsilon(s)), \text{ with } f(x, \epsilon) := \mathrm{i}k(-x^2 + \epsilon) \tag{*}$$

and the *control function* $\varepsilon(\cdot)$.

The solution $x(\cdot)$ blows-up in the time-points s such that $y(s) = 0$. The simplest way to describe the evolution of x near ∞ is to see that, when x evolves in the neighborhood $\widehat{\mathbb{C}} \setminus \{0\}$ of ∞, the dynamics of

$$\widetilde{x}(s) = -1/x(s)$$

satisfies

$$\widetilde{x}' = \widetilde{f}(\widetilde{x}, \varepsilon), \qquad \text{where } \widetilde{f}(\widetilde{x}, \epsilon) := \mathrm{i}k(-1 + \epsilon\widetilde{x}^2).$$

Recall that a state $x(s_-)$ of the system (*) is said to be in the *time-t-controllable set* $\mathcal{C}_{\{t\}}(\eta_+, k)$ to a state η_+ with $t \geq 0$ if there exists a feasible control $\varepsilon \in \mathbb{F}_{s_-}$ such that $x(s_- + t) = \eta_+$.

For $t \in [0, +\infty]$, we put

$$\mathcal{C}_{[0,t)}(\eta_+, k) := \bigcup_{0 \leq t_0 < t} \mathcal{C}_{\{t_0\}}(\eta_+, k).$$

A state η_- is said to be *controllable to* η_+ if it belongs to the set $\mathcal{C}_{[0,+\infty)}(\eta_+)$.

A feasible control $\varepsilon \in \mathbb{F}_{s_-}$ is said to be a *minimum-time control* from $x(s_-)$ to η_+ if $x(s_- + t) = \eta_+$ in the minimum possible time $t = T_k^{\min}(x(s_-), \eta_+)$, which can be defined by

$$T_k^{\min}(x(s_-), \eta_+) := \inf\{t \geq 0 \ : \ x(s_-) \in \mathcal{C}_{\{t\}}(\eta_+, k)\}.$$

If η_- is not controllable to η_+, we put by definition $T_k^{\min}(\eta_-, \eta_+) := +\infty$.

Definition. For $(\eta_-; \eta_+) \in \widehat{\mathbb{C}}^2$, we say that $k \in \mathbb{C} \setminus \{0\}$ is an $(\eta_-; \eta_+)$-eigenvalue of $\varepsilon(\cdot)$ on an interval (s_-, s_+) if equation (Eq) has a nontrivial solution y satisfying the two boundary conditions

$$\frac{y'(s_\pm)}{\mathrm{i}k y(s_\pm)} = \eta_\pm \quad \text{(which are understood as } y(s_\pm) = 0 \text{ when } \eta_\pm = \infty).$$

We denote the set of $(\eta_-; \eta_+)$-eigenvalues by $\Sigma_{\eta_-, \eta_+}^{s_-, s_+}(\varepsilon)$.

Now the following proposition is obvious.

Proposition. *The following statements are equivalent for $t > 0$:*

(C1) $k \in \Sigma_{\eta_-, \eta_+}^{s_-, s_+}(\varepsilon)$, where $s_+ = s_- + t$;

(C2) $k \neq 0$ and the control $\varepsilon(\cdot)$ steers the system (*) from the initial state η_- to the target state η_+ in time t.

(CY) $k \neq 0$ and $\varepsilon(\cdot)$ steers the system (YEq) from $Y(s_-) = Y^{[\eta_-]}$ to the state $Y(s_- + t)$ that belong to the target plane $\mathfrak{T}^{[\eta_+]}$, where

$$Y^{[\eta_-]} := \begin{cases} (1; \eta_-)^\top, & \text{if } \eta_- \neq \infty; \\ (0; 1)^\top, & \text{if } \eta_- = \infty; \end{cases}$$
$$\text{and} \qquad \mathfrak{T}^{[\eta_+]} := Y^{[\eta_+]}\mathbb{C} = \{a Y^{[\eta_+]} \ : \ a \in \mathbb{C}\}.$$

In particular, the problem (MinL) is equivalent to the problem of minimum-time control of the system (*) from $(-\mathfrak{n}_\infty)$ to $\mathfrak{n}_\infty$.

Remark. Assume that $s_+ = -s_- = d > 0$. Then

$$\Sigma^{\mathrm{odd}}(\varepsilon) = \Sigma_{-\mathfrak{n}_\infty, \infty}^{-d, 0}(\varepsilon) = \Sigma_{\infty, \mathfrak{n}_\infty}^{0, d}(\varepsilon)$$

and

$$\Sigma^{\mathrm{even}}(\varepsilon) = \Sigma_{-\mathfrak{n}_\infty, 0}^{-d, 0}(\varepsilon) = \Sigma_{0, \mathfrak{n}_\infty}^{0, d}(\varepsilon)$$

for $\varepsilon \in \mathbb{F}_d^{\mathrm{sym}}$. One can formulate an analogue of problem (MinL) for odd-mode (even-mode) resonances, and then reformulate them as minimum-time control problems, for example, from $(-\mathfrak{n}_\infty)$ to ∞ (resp., from $(-\mathfrak{n}_\infty)$ to 0).

12.3 The connection of dual problem with the original problem of Pareto optimization

Let the interval $[s_-, s_+]$ be again fixed.

We assume in this section that

$$\mathfrak{n}_1 \le \mathfrak{n}_\infty \le \mathfrak{n}_2.$$

(i.e., the permittivity $\epsilon_\infty = \mathfrak{n}_\infty^2$ of the outer medium is in the range of permittivities available for the fabrication process).

Theorem. *Assume stronger inequalities $\mathfrak{n}_1 < \mathfrak{n}_\infty < \mathfrak{n}_2$. Assume that k is the resonance of minimal decay for a frequency $\alpha = \operatorname{Re} k$. Then the family of minimum time controls for (*) steering $x(s_-) = (-\mathfrak{n}_\infty)$ to $\mathfrak{n}_\infty$ coincides with the family of resonators of minimal decay for the frequency α. In particular,*

$$T_k^{\min}(-\mathfrak{n}_\infty, \mathfrak{n}_\infty) = s_+ - s_-.$$

Remark. In the cases $\mathfrak{n}_\infty = \mathfrak{n}_1$ or $\mathfrak{n}_\infty = \mathfrak{n}_2$, the reduction is "partial" in the sense that at least some of resonators of minimal decay can be constructed from each of minimum-time controls.

The rest of this section gives one of the main steps of the proof. of the above theorem.

Let $\operatorname{Arg}_0 z$ is a continuous in $z \in \mathbb{C} \setminus \overline{\mathbb{R}_-}$ branch of the multi-valued complex argument $\operatorname{Arg} z$ fixed by $\operatorname{Arg}_0 1 = 0$.

Let us consider the problem of minimization of modulus $|k|$ of an $(\eta_-; \eta_+)$-eigenvalue k for a given complex argument $\gamma = \operatorname{Arg}_0 k$ over the family $\mathbb{F}_{s_-,s_+}$, where the finite interval (s_-, s_+) with $s_- < s_+$ and the tuple $(\eta_-; \eta_+)$ are fixed.

The main tool for this reformulation is the natural scaling of the eigenproblem:

$$\text{if } k \in \Sigma_{\eta_-,\eta_+}^{s_-,s_+}(\varepsilon) \text{ and } \widetilde{\varepsilon}(s) = \varepsilon(\tau s) \text{ for } \tau \in \mathbb{R}_+, \text{ then } \tau k \in \Sigma_{\eta_-,\eta_+}^{\tau^{-1}s_-,\tau^{-1}s_+}(\widetilde{\varepsilon}).$$

Let us introduce the set

$$\Sigma_{\eta_-,\eta_+}^{s_-,s_+}[\mathbb{F}_{s_-}] := \bigcup_{\varepsilon \in \mathbb{F}_{s_-,s_+}} \Sigma_{\eta_-,\eta_+}^{s_-,s_+}(\varepsilon)$$

of *achievable* $(\eta_-; \eta_+)$*-eigenvalues* (over $\mathbb{F}_{s_-,s_+}$).

We define the *minimal modulus* $\rho_{\min}(\gamma) = \rho_{\min}(\gamma, \eta_-, \eta_+)$ by

$$\rho_{\min}(\gamma) := \inf\{|k| \ : \ k \in \Sigma_{\eta_-,\eta_+}^{s_-,s_+}[\mathbb{F}_{s_-}] \text{ and } \operatorname{Arg}_0 k = \gamma\}. \qquad (**)$$

and the set of *achievable* $(\eta_-; \eta_+)$*-arguments* by

$$\operatorname{dom} \rho_{\min} := \{\operatorname{Arg}_0 k \ : \ k \in \Sigma_{\eta_-,\eta_+}^{s_-,s_+}(\varepsilon) \text{ for certain } \varepsilon \in \mathbb{F}_{s_-,s_+}\}.$$

The function $\rho_{\min}$ takes values in $[0, +\infty]$ and depends on γ, $\eta_\pm$, and $s_+ - s_-$. We omit $s_\pm$ and sometimes $\eta_\pm$ from the list of variables of $\rho_{\min}$ when they are fixed.

If $k_\gamma^{\min} := \mathrm{e}^{\mathrm{i}\gamma}\rho_{\min}(\gamma)$ belongs to $\Sigma_{s_-,s_+}^{\eta_-,\eta_+}(\varepsilon_\gamma^{\min})$ for a certain $\varepsilon_\gamma^{\min}(\cdot) \in \mathbb{F}_{s_-,s_+}$, i.e., if minimum is achieved in (**), then we say that

$$\varepsilon_\gamma^{\min}(\cdot) \text{ is a resonator of minimal modulus } |k| \text{ for (the complex argument) } \gamma.$$

The set $P_{\mathrm{mod}}^{\eta_-,\eta_+} := \{\mathrm{e}^{\mathrm{i}\gamma}\rho_{\min}(\gamma) \ : \ \gamma \in \operatorname{Arg}_0 \Sigma_{\eta_-,\eta_+}^{s_-,s_+}[\mathbb{F}_{s_-,s_+}]\}$ forms the *Pareto optimal frontier* for the problem of minimization of the modulus $|k|$ of an $(\eta_-;\eta_+)$-eigenvalue k over $\mathbb{F}_{s_-,s_+}$.

The minimum-time control problem for the system (*) and the problem of finding of resonators of minimal modulus for given γ over $\mathbb{F}_{s_-,s_+}$ are equivalent in the sense of the following theorem, which includes also a result on the existence of optimizers.

Theorem. *Let $\eta_- \neq \eta_+$, $k \neq 0$, and $\gamma = \operatorname{Arg}_0 k$. Then the following statements are equivalent:*

(i) $\eta_- \in \mathcal{C}_{[0,+\infty)}(\eta_+, k)$, i.e., (*) is controllable from η_- to η_+;

(ii) there exists a minimum-time control $\varepsilon(\cdot) \in \mathbb{F}_{s_-}$ for (*) that steers η_- to η_+ in the minimal time $T_k^{\min}(\eta_-, \eta_+)$;

(iii) $\gamma \in \operatorname{dom} \rho_{\min}(\cdot, \eta_-, \eta_+)$;

(iv) there exist at least one resonator $\varepsilon_\gamma^{\min}(\cdot)$ of minimal modulus for γ over $\mathbb{F}_{s_-,s_+}$.

If statements (i)-(iv) hold true, then

$$T_k^{\min}(\eta_-, \eta_+) = \frac{(s_+ - s_-)\rho_{\min}(\gamma, \eta_-, \eta_+)}{|k|}. \tag{5}$$

If, additionally, $s_\pm$ are chosen so that $s_+ - s_- = T_k^{\min}(-\eta_-, \eta_+)$, then the families of minimum-time controls $\varepsilon(\cdot)$ and of resonators of minimal modulus $\varepsilon_\gamma^{\min}(\cdot)$ coincide.

Proposition. *Let $\mathfrak{n}_1 \leq \mathfrak{n}_\infty \leq \mathfrak{n}_2$ and $\eta_\pm = \pm\mathfrak{n}_\infty$. Then*

$$\Sigma[\mathbb{F}_{s_-,s_+}] = \{ce^{\mathrm{i}\gamma}\rho_{\min}(\gamma, -\mathfrak{n}_\infty, \mathfrak{n}_\infty) \ : \ c \in [1, +\infty) \text{ and } \gamma \text{ is achievable } \}.$$

(ii) The Pareto frontier P of minimal decay can be found from the Pareto frontier $P_{\mathrm{mod}}^{-\mathfrak{n}_\infty,\mathfrak{n}_\infty}$ of minimal modulus.

The last proposition follows from Lemma in subsection 12.1.

References for Section 12.

[KKV18] I.M. Karabash, H. Koch, I.V. Verbytskyi, Pareto optimization of resonances and minimum-time control, preprint arXiv:1808.09186, https://arxiv.org/pdf/1808.09186